Cognitive Component Analysis

Feng, Ling

Publication date:
2008

Document Version
Publisher's PDF, also known as Version of record

Cognitive Component Analysis

Ling Feng

Kongens Lyngby 2008
IMM-PHD-2008-196

Technical University of Denmark
Department of Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-PHD: ISSN 0909-3192

Summary

This dissertation concerns the investigation of the consistency of statistical regularities in a signaling ecology and human cognition, while inferring appropriate actions for a speech-based perceptual task. It is based on unsupervised Independent Component Analysis providing a rich spectrum of audio contexts along with pattern recognition methods to map components to known contexts. It also involves looking for the right representations for auditory inputs, i.e. the data analytic processing pipelines invoked by human brains.
The main ideas refer to Cognitive Component Analysis, defined as the process of unsupervised grouping of generic data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity. Its hypothesis runs ecologically: features which are essentially independent in a context defined ensemble, can be efficiently coded as sparse independent component representations.
The focus has been to construct a preprocessing pipeline for COCA to search for the 'cognitive structure', and to measure the alignment of the resulting from unsupervised learning and human cognition. Based on the nature of human auditory system and psychoacoustics, we have constructed the pipeline: feature extraction; feature integration; energy based sparsification; and principal component analysis. To test whether human uses information theoretically optimal ICA methods in higher cognitive functions, is the main concern in this thesis. It is well-documented that unsupervised learning discovers statistical regularities. However human cognition is too complicated and not yet fully understood. Nevertheless, in our approach we represent human cognitive processes as a classification rule in supervised learning. Thus we have devised a testable protocol to test the consistency of statistical properties and human cognitive activity, i.e. unsupervised learning of perceptual inputs and supervised learning of inputs together with manually obtained labels. The comparison has been carried out at different levels. This protocol has successfully revealed the consistency of two classifications via several speech-based cognitive tasks.

Resumé

Denne afhandling undersøger sammenhængen mellem talesignalers statistiske egenskaber og kognitive processer involveret i taleforståelse. Der tages udgangspunkt i unsupervised Independent Component Analysis og der analyseres et bredt spektrum af lyd-sammenhænge og relevante metoder til mønstergenkendelse. Afhandlingen søger også at finde repræsentationer for auditive inputs, det vil sige strukturen i den databehandling som bruges i menneskets hjerne.
Hovedbidraget vedrører såkaldt Cognitive Component Analysis (COCA), defineret som en proces af 'unsupervised' opdeling af generiske data, således at den fundne struktur ligger sig tæt op ad strukturen fra menneskets kognitive aktiviteter. Hypotesen er følgende: features, som grundlæggende er uafhængige i en given sammenhæng, kan effektivt kodes som repræsentationer af 'sparse independent components'.
Fokus har været på at konstruere en data-analytisk 'pipeline' til COCA for at finde den kognitive struktur, og måle hvorledes unsupervised opdeling stemmer overens med den menneskelige kognitions måde at gruppere indtryk på. Baseret på den menneskelige hørelse og psykoakustiske principper har vi konstrueret en pipeline der består af følgende trin: feature udtrækning; feature integration; energi-baseret sparsification; og principal komponent analyse. At teste om mennesker bruger informations teoretisk optimale ICA metoder i de højere kognitive funktioner er hovedformålet med denne afhandling. Det er grundigt dokumenteret at unsupervised learning afslører statistiske regulariteter. Den menneskelige kognition er derimod meget kompliceret og endnu ikke fuldstændigt forstået. I vores fremgangsmåde forsøger vi at repræsentere menneskets kognitive mekanismer som klassifikationsregler i supervised learning. Dermed har vi konstrueret en testbar protokol, der kan bruges til at evaluere sammenhængen mellem statistiske egenskaber og menneskets kognitive aktivitet, det vil sige, sammenhængen mellem unsupervised learning af perceptuelle input og supervised learning af input sammen med manuelle klassifikationer. Sammenligningen er blevet gennemført på forskellige niveauer. Denne protokol har med succes påvist denne sammenhæng i to klassifikationsopgaver.

Preface

This dissertation was prepared at the Department of Informatics Mathematical Modelling, the Technical University of Denmark in partial fulfillment of the requirements for acquiring the Ph.D. degree in electrical engineering.

The dissertation mainly covers the research on Cognitive Component Analysis (COCA), which was proposed to examine the hypothesis that human cognition uses information theoretically optimal ICA methods for generic data analysis. The project focused on speech, including the mathematical modeling of speech signals, and the comparison between unsupervised modeling of data and that resulting from human cognition. As a subproject, music was of interest. To provide a unifying framework of COCA, the music study is only covered in appendix I.

This dissertation includes a summary report and a collection of eight research papers written during the period 2005–2008, and elsewhere published.

Lyngby, April 2008

Ling Feng

Contributions

This project has produced one journal article, seven conference papers, several abstracts and technical reports. All papers can be found at http://orbit.dtu.dk/app and http://www.imm.dtu.dk/English/Research/ISP/Publications.aspx. The main contributions are included in Appendix.

The work about vocal segment classification in music was inspired during the external research stay in Lee-lab, the Institute for Neural Computation (INC), University of California, San Diego (UCSD), and has been continued in DTU Informatics. Appendix I summarizes this work.

Journal Paper

Ling Feng, Lars Kai Hansen. Cognitive Components of Speech. *Artificial Intelligence Journal*, 2008. Submitted. Appendix [H].

Conference Papers

Ling Feng, Lars Kai Hansen. On Low-level Cognitive Components of Speech. *International Conference on Computational Intelligence for Modelling* (CIMCA), vol.2, pp 852-857, 2005. Published. Appendix [B].

Ling Feng, Lars Kai Hansen. Phonemes as Short Time Cognitive Components. *The 31st International Conference on Acoustics, Speech, and Signal Processing* (ICASSP), vol.5, pp 869-872, 2006. Published. Appendix [C].

Lars Kai Hansen, Ling Feng. Cogito Componentiter Ergo Sum *6th International Conference on Independent Component Analysis and Blind Source Separation* (ICA), pp 446-453, 2006. Published. Appendix [D].

Ling Feng, Lars Kai Hansen. Cognitive Components of Speech at Different Time Scales. *The 29th annual meeting of the Cognitive Science Society* (CogSci), pp 983-988, 2007. Published. Appendix [E].

Ling Feng, Lars Kai Hansen. On Phonemes as Cognitive Components of Speech. *The 1st IAPR Workshop on Cognitive Information Processing* (CIP), pp 205-210, 2008. Published. Appendix [F].

Ling Feng, Lars Kai Hansen. Is Cognitive Activity of Speech Based on Statistical Independence? *The 30th annual meeting of the Cognitive Science Society* (CogSci), pp 1197-1202, 2008. Published. Appendix [G].

Ling Feng, Andreas Brinch Nielsen, Lars Kai Hansen. Vocal Segment Classification in Popular Music. *The International Conferences on Music Information Retrieval and Related Activities* (ISMIR), pp 121-126, 2008. Published. Appendix [I].

Abstracts

1 Ling Feng, Cognitive Component Analysis. *Machine Learning: Theory, Applications, Experiences. A Workshop for Women in Machine Learning*, 2006.

2 Ling Feng, Lars Kai Hansen. Cognitive Components of Speech at Different Time Scales *NIPS Workshop: Advances in Acoustic Models*, 2006.

Acknowledgements

I would like to thank Professor Lars Kai Hansen for providing me with such a good opportunity for a Ph.D. study in Intelligent Signal Processing section, DTU Informatics. Prof. Hansen's constructive supervision has inspired me throughout the whole Ph.D study, and his wide experience and extensive knowledge have guided me to many interesting and attractive research areas. I also thank the staff in ISP group for pleasant company, especially Andreas Brinch Nielsen for his collaboration and assistance. Warm thanks go to the people who spent their precious time on helping me construct the music database. Special thank goes to the department secretary Ulla Nørhave for being concerned and helpful all the time.

I would also like to thank Professor Te-Won Lee and his Lee-Lab members for having me during my external research study in UCSD. I appreciate their hospitality and collaboration, and feel privileged to have worked with Lee-lab, especially Jiucang Hao, who has never been stingy on spending time for discussions. His help made my stay pleasant and free from worries.

Furthermore I am grateful to the Danish Technical Research Council for supporting 'Intelligent Sound' project, which also made my Ph. D. project possible. As well I give thanks to Otto Mønsteds Fond, Reinholdt W. Jorck og Hustrus Fond, Marie & M.B. Richters Fond, Oticon Fonden, and Niels Bohr Legatet for financial support, which financially supported my six-month external stay and my attendance of several international conferences and workshops.

Last but not the least, I thank my parents and family for their love, and thank Erling for his proofreading of my papers, and more importantly for his company in the late nights of my final thesis writing period.

Contents

Chapter 1

Introduction

Human cognition involves the mental process of knowing, and it is the action and interaction of multiple brain functions, ranging from perception and construction, calculation ability, attention (information processing), memory, to executive functions such as planning, problem-solving, and self-monitoring. The scientific exploration of human cognition is being transformed by novel techniques for representing activities of the intact human brain, while neurobiological sources are prevailing inputs to cognitive relevant processing structures. This transformation will indeed lead to some promising breakthroughs in our scientific understanding of human capacities. The natural cognitive system by then will likely be provided with a more detailed analysis in the neuroscience viewpoint. However to be able to build detailed computational models of human cognition is the foundation.

Let us take the speech understanding as an example. Speech perception is one of the fundamental cognitive behaviors of humans. The study of statistical modeling of speech has been carried out for more than two decades, and the resulting state-of-the-art automatic speech recognition (ASR) system, commonly with HMM-based processes, has reached its plateau. The independent development of ASR does not take many mechanisms of human speech perception into account, and the outperforming of human speech perception is still quite obvious in many situations, such as noisy environments, conversational speech and spontaneous speech, etc. Therefore the demand of detailed computational

models which are able to adopt and account for how human speech understanding works, leads to a future research tend [19]. This step will benefit two communities: both the ASR research community and human speech perception community, and eventually make a significant contribution to closing the gap between machines performance and human cognition.

This opens a question about the relation between machine processes and human processes (neuro-biological processes). What interests us is the issue: *Whether the characteristics of human processors are determined by statistical properties of the perceived information?* The positive proof of this question, and moreover the close relation between statistical regularities and characteristics of human speech processing, will allow us to further ponder on how to improve the situation. A natural query one can pose could be: 'How to adapt the current ASR system to these statistical regularities for system modification?', which will guide us to implement and build new statistical models to imitate human processes.

This dissertation is mainly searching for an answer of the fundamental question: *Do the characteristics of human processors reflect statistical regularities revealed by unsupervised learning of perceptual inputs?* Consequently Cognitive Component Analysis (COCA) has been proposed as the process of unsupervised grouping of data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity [30]. It aims at investigating the consistency of statistical regularities in a signaling ecology and human cognitive activity.

1.1 The Course of Cognitive Component Analysis

With the ongoing research on COCA, it has been progressed from a tentative assumption to a quite well developed analysis. Inspired by the success of independent component analysis (ICA) in relevant natural ensemble statistics, one of our hypotheses is based on independence. In the beginning of this project, some low-level COCA has been touched upon on speech signals using unsupervised linear component analysis algorithms. In appendix B, principal component analysis (PCA) has been carried out on speech data to reveal 'fingerprints' stemmed from phonemes and speaker identities. Appendix C has proved the significance of introducing ICA on COCA to discover phoneme signatures. The generality of COCA has been shown on many topics in appendix D.

The theoretical background of COCA has been gradually built up based both on cognitive psychology, e.g. the evolution of human brain, and on computer

science, e.g. variance mathematical models which try to offer human-like functions. The sole pursue of comparable performance of machines to humans by optimizing modeling algorithms, has been diverted to a thorough study of compatibility and difference between machines and humans performance at every level, which has become the new research tide for pursuing a possible solution to natural human-computer interfaces. It comes the time to further impel our COCA to a comparison level based on models classification results. Statistical regularities can be discovered by unsupervised learning methods, furthermore supervised learning of manual labels loosely represents human cognitive activity. Therefore, the comparison between unsupervised and supervised learning could be one such method to test the influence of statistical regularities on human cognition. The modified version of mixture of factor analysis (MFA) is a candidate classifier. Appendix E has carried out this idea. MFA has been modified into ICA-like density models for both unsupervised learning and supervised learning, and the only difference is the modeling of human labels in supervised learning MFA model.

While MFA is capable of proving the dependency of human cognition on statistical regularities, it did not reflect the statistical independence in a straightforward manner. A more flexible set of models have been designed. In appendix F, G and H, ICA+naive Bayes model has been compared with the mixture of Gaussians (MoG) on speech signals involving phonemes, gender, age and speaker identity, at three levels based on classification performance. High correlation of both models has once again proved that human cognitive activity is based on the statistical independence as one of the statistical regularities.

1.2 Dissertation Reading Guide

This dissertation follows the itinerary of searching for spoken cognitive components, and answers for the fundamental questions. It is organized as follows:

Chapter 2: COCA is based on ecology, physiology and machine learning. To fully describe our intention, we open this dissertation with human cognition. The emphasis is allocated on brain structures from the anatomical point of view, by which means will help us understand brain functions. Since cognitive components of speech are our main interests, to understand how human ear perceive and process sound becomes influential. Human auditory system will be introduced, and it gives us the physiological background and the concrete base for building COCA preprocessing pipeline.

Chapter 3 gives the overview of cognitive component analysis, including the

theoretical background, motivation, hypothesis, and basic processing scheme. Some simple examples on revealing cognitive components of various topics show the generality of COCA analysis. They are beyond the scope of this dissertation, nevertheless serve as appetizers.

Chapter 4 describes the low-level cognitive component analysis. The chapter starts with some basic introduction of machine learning. A general unsupervised hidden variable model is discussed, since it is the basic formulation for many models, which are involved in COCA. Examples demonstrate the ability of some unsupervised learning models in unveiling 'ray structure', in other words cognitive components. In the studies of phonemes, we are able to discover 'invariant cue' in different conditions by applying unsupervised grouping of data.

Chapter 5 investigates the higher level cognitive component analysis. Feature integration constructs data at different time scales, and we analysis data with attempt to discover higher cognitive functions. It introduces the devised protocol for testing cognitive consistency: unsupervised vs. supervised learning. Two sets of models will be introduced in a row. One is the unsupervised and supervised version of modified MFA based on ICA-like density models. As the inspiration of these models, *Soft-LOST* and *Hard-LOST* models will be discussed. The second set is the ICA+naive Bayes vs. MoG model. The experimental design for model comparison follows. Whereafter comparison methods at several levels are described.

Chapter 6 recapitulates the main ideas of cognitive component analysis of speech. The conclusion will run through the development of COCA, and summarize findings from each step of the research. A few perspectives will shed some light on the future work.

Chapter 2

Human Cognition

This chapter gives a brief introduction about the basic knowledge of human cognition, where perception is one of the significant shares. From the anatomical point of view, we elucidate brain structures, so as to refer to the corresponding brain functions. Cognitive component analysis covers many topics, from text mining, music to social network, etc. Nevertheless this dissertation basically describes the cognitive component analysis of speech signals, and it brings the necessariness to understand human audio perception and the functionality of human ears. Human cognition is complex and sophisticated, and not yet fully discovered and understood. The brief introduction given in the chapter will try to go through the basic, and devote more space to those which are best understood.

Cognition, in Latin *cognoscere* meaning 'to know', represents the human or human-like processing of information using knowledge and preferences. It often refers to mental functions, mental processes and states of intelligent entities, such as humans and highly autonomous robots. Such mental processes essentially include learning, comprehension, inferencing, planning, decision-making, judging and problem-solving, etc. Therefore the concept of cognition is closely related to such abstract concepts as mind, reasoning, perception, intelligence and those that describe numerous capabilities of the human mind. The definition of cognition is not yet concrete, a broader employ of cognition is referred to as the act of knowing in a social or cultural viewpoint, to depict the development

of concepts and knowledge within a group that culminates in both thought and action. [51].

Human cognition usually considered the same as cognitive modeling and cognitive architecture, is the study of how human brains work. Some models are built based on the understanding of neurobiology, e.g. the characteristics of neurons and their connectivity in the brain. The model performance in demonstrating human-like behavior is the main evaluation criteria. Artificial intelligence as one of the prevailing research area, is also defined with reference to human cognition. Cognitive process is the result of the interplay between statistical properties of the ecology and the process of natural selection along with human evolution. Human brains can learn information from lower levels, where they have been posited. Furthermore they can integrate information from experiences as well, to build higher levels of meaning and cognition. In other words, learning and inference functions of the brain enable us to infer the proper action to a given situation, even this situation has not yet been experienced by us. The evolution optimizes human brain, and the resulting human cognitive system can model complex multi-agent scenery, and use a broad spectrum of cues for analyzing perceptual input, and for identification of individual signal processes.

Learning defined as the acquisition and development of memories and behaviors, is the product of experiences. The perception of the world provides us with the abilities to observe what happens in the world everyday, by hearing, watching, touching, tasting and smelling, etc. In the next section, we will introduce the concept of perception, and put more effort introducing relevant brain structures and functions.

2.1 Perception and Brain Functions

Perception is originated from a Latin word '*percepio*'. It means receiving, collecting, action of taking possession, apprehension with the mind or senses. To human beings, the perception of the world seems so natural, straightforward, immediate, effortless, and nearly accurate from an intact brain. However it involves a huge mass of neurons to carry out complex operations in the brain, and the cerebral cortex, as the most highly developed structure of a brain, is dedicated largely to perception. The perceptual input is divided into five major groups, shown in Table 2.1, which is based on the classification of senses [53]. The nervous system transmits information by means of electrical signals, called neural impulses, passing from one cell to another. Such neural impulses are stimulated by a number of forms of environmental energy, e.g. the sound energy as the form of air pressure waves are sensed by ears. There are specialized

Table 2.1: The five senses

Sense	Receptor	Sensory structure	Cortex
Vision	Photoreceptors	Eye	Visual cortex
Hearing	Mechanoreceptors	Ear	Auditory cortex
Touch	Mechanoreceptors, Thermoreceptors	Skin, Muscle, etc.	Somatosensory cortex
Balance	Mechanoreceptors	Vestibular organs	Temporal cortex
Taste Smell	Chemoreceptors	Nose, Mouth	Primary taste cortex, Olfactory cortex

cells serving each sense, meaning that they respond to one particular form of energy and convert it to neural signals. As listed in Table 2.1, a particular sensory structure or organ senses the corresponding energy form, and the energy is received by some particular receptors. However neural impulses generated by different sensory organs have the same form, and it is difficult to tell which impulse is generated by which organ, by looking at the signal itself. The difference lies in the location where this signal's transportation ends in the brain, such as the information sensed by the ear will be conveyed to the auditory cortex [53].

For more than a century we have found out that different parts of the brain have different functions. The clinical evidences prove that symptoms are heavily related to the locations of brain damages. Therefore we need to understand brain structures and their functions. The following introduction is based on [51] and [53]. The human brain consists of hindbrain (also known as the prosencephalon), midbrain (mesencephalon) and forebrain (rhombencephalon). The midbrain is entirely hidden under the forebrain, which also covers most of the hindbrain. The hindbrain is composed by three parts: *medulla oblongata*, *pons* and *cerebellum*. Each has its own functions. *Cerebellum* is the only part of hindbrain visible in Figure 2.1. Loosely speaking, *medulla oblongata* controls the heartbeat, and life critical functions as such. *Pons* acts as a relay station, which conveys signals from various parts of the cerebral cortex to *cerebellum*, and it also regulates breathing. As it is commonly known, *cerebellum* as the largest part of the hindbrain, stores the learnt movement, and consequently coordinates the body balance and movements. Brain damages in this area will cause rough and jerky movements. Recent study has suggested that the cerebellum has also the functioning of spatial reasoning, sound discriminating, etc [11].

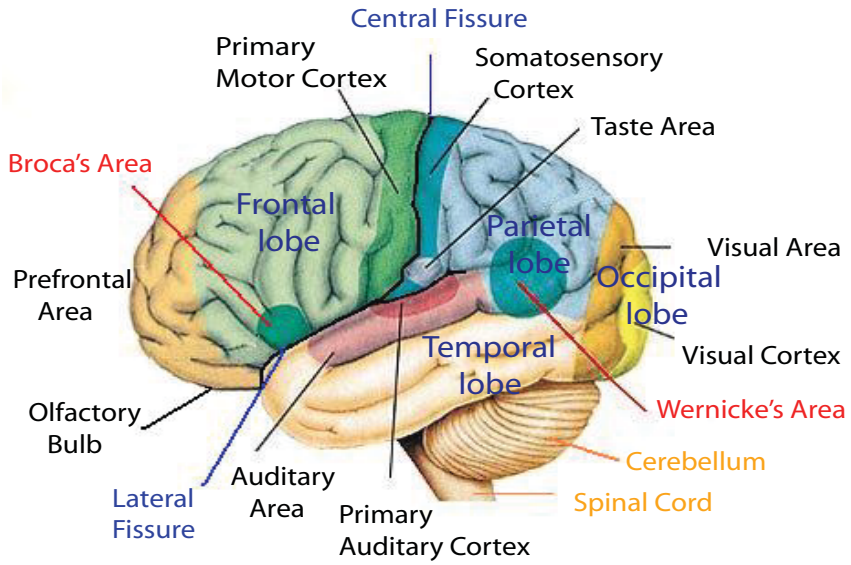

Figure 2.1: Human cerebral cortex. The view shows the outer surface of the forebrain of the brains' left cerebral hemisphere. The left-hand side is the front of the head. The structure of the right cerebral hemisphere is similar, even they have somewhat different functions. The four lobes: frontal, parietal, temporal and occipital lobes are separated by fissures. Major functions of various parts of lobes have been shown, corresponding to cortexes of the five senses in Table 2.1.

The midbrain helps us perceive the world with vision and hearing. It controls movements, especially eye movements, and also transmits auditory information received from ears to the auditory area of the forebrain. The forebrain is the most interesting part of the brain, since its outer surface, the cerebral cortex, is involved in various functions. Figure 2.1 gives a picture of human cerebral cortex.

As the extensive outer layer of 'gray matter' of cerebral hemispheres, the cerebral cortex comprises about 80% of a human brain. It is approximately $2.5mm$ thick, and $1000cm^2$ in surface area if stretched out flat. However the cortex is all crumpled up and squeezed into a limited space inside the skull, which gives the wrinkle looked surface of the forebrain. Some parts of the cortex reach deep inside the brain, which forms grooves separating the brain into different sections. The deepest groove divides the brain into left and right cerebral hemisphere, and it is called longitudinal fissure. The corpus callosum connects the left hemisphere to the right. Due to the crossing over of the spinal tracts, the left hemisphere

deals with the right side of the body, and vice versa. On each side of the brain, grooves further divide the cortex into four lobes: frontal, parietal, temporal and occipital lobe, see Figure 2.1. The frontal lobe and parietal lobe are separated by the central fissure at the top of the brain, and the lateral fissure differentiates the bottom of the frontal lobe and the temporal lobe. The primary motor cortex has also influence on movements, and the somatosensory cortex senses experiences as touch and various feelings it may rise. Besides the mentioned functionalities, the cerebral cortex also provides us with capabilities of speaking and language comprehension; higher thought processes, such as logic and reasoning, planning; memory, personality and other human activities. Now let us get familiar with brain functions by delving into particular functionalities of these four lobes on the forebrain.

- Frontal lobe is responsible for motor activities and the integration of muscle activities steered by the motor area; speech ability controlled by the *Broca's area*; thought processes, like planning, concentration, emotional traits, judgment and inhibition, and so on, dominated by the prefrontal area.
- Parietal lobe is associated with the sensory cortex to process sensory input and discriminate sensory information, e.g. touch, taste, pressure, pain, heat and cold. It is believed that this area is also responsible for reading and arithmetic.
- Temporal lobe mainly receives auditory signals, through primary auditory cortex and secondary auditory cortex. Furthermore an area called *Wernicke's area* gives temporal lobe a function of language comprehension. More details will be given later. The temporal lobe is also responsible of making new memories, and serves as an emotion evaluator. Patients suffering **Capgras Syndrome** is believed to have their **amygdala** injured, which locates in the temporal lobe [21].
- Occipital lobe is the smallest one of the four lobes. Locating in the rearmost part of the skull, it serves as the visual processing center. It receives information from eyes, then processes and associates the information with images stored in memory for discrimination of movements and color recognition, etc.

The emergence of languages is a milestone of the human evolution. It not only helps us communicate, but build a structural representation in our mind as well. As mentioned briefly, *Wernicke's area* is responsible for the language processing, comprehension, and the interpreting of words; and *Broca's area* is associated with speaking ability. The former one encircles the auditory cortex, where the

temporal lobe and parietal lobe meet. For most people, both of these areas locate on the left cerebral hemisphere of the brain, which brought the common acknowledgement that human speech and language ability are steered by the left hemisphere. Two areas are connected via a neural pathway: the arcuate fasciculus. The lesion on *Wernicke's area* will influence the understanding of language, and the understanding of written and spoken words; and patients with lesion on *Broca's area* are unable to create grammatically-complex sentences. The study of human language processing has shown that brain uses three types of interacting structures to perform speech processing:

1. the concepts structure involving non-linguistic missions;
2. the linguistic structure involving the mental lexicon and syntactic rules;
3. the mediation structure working as an interconnection between 1 and 2.

Not all these structures locate in the left hemisphere: the conceptual related structure sits in both cerebral hemispheres; and the latter two mostly locate in the left hemisphere [19].

As our interests lie in the area of speech perception and cognitive behaviors related to speech, how the human ear processes speech or, in general, sounds is the base of our study. The next section will elaborate on the physiology of human auditory system.

2.2 Physiology of Human Auditory System

Locating at the back of the brain, the occipital lobe is well protected from injuries, and serves humans to explore the world from visual inputs. Humans are considered as highly visually advanced animals. However the immediate field of view is limited to merely 200^o, and visual perception also relies on the brightness of environments. On the contrary, our auditory perception is more relaxed. The sound perception happens all the time consciously or unconsciously, and there is no such tissue like eyelid, which can switch on and off the visual perception as you prefer.

The only directly visible part of the human peripheral auditory system, the flexible flap surrounding the outer ear, sits on both sides of the head. Human auditory system can be roughly seen as the combination of two systems: the peripheral auditory system and the central auditory system. The latter includes

a mass amount of neurons in the brainstem and the cerebral cortex. Since the understanding of the peripheral auditory system is more thoroughgoing than the central auditory system, here the physiology of auditory system will be recounted. Further, to design features which are capable of emulating the way human auditory system processing sounds, we need to have a close look at internal components of the human ear as well.

The peripheral auditory system consists of three parts: the outer ear, the middle ear and the inner ear. Detailed components are shown in Figure 2.2.

2.2.1 The Outer Ear

The outer ear includes pinna and ear canal. Pinna begins with the flexible flap, and ends with the funnel-shaped inner part, known as the concha. The shape of the pinna reflects and diffracts sounds, and the processed signal contains the information for sound localization. Moreover folds of the pinna attenuate high-frequency sound components, and this behavior can be regarded as a filter.

The ear canal is about $25mm$ long and $7mm$ in diameter. It can be described as a slightly bended tube, and is closed with semi-transparent membrane, namely tympanic membrane. This membrane is tilted to form a 60 degrees angle with reference to the ear canal's axis, which gives the membrane a cone shape inwards the middle ear. The overall shape of the outer ear works as an amplifier.

2.2.2 The Middle Ear

The middle ear transmits the vibration of the tympanic membrane to movement of the fluid in the inner ear. This transmission is carried out by three interconnected bones: *malleus*, *incus* and *stapes*. They are named in Latin, meaning hammer, anvil and stirrup separately. The hammer is fixed to the inner surface of the tympanic membrane, and the anvil connects the hammer and stirrup. The smallest bone in the body, the stirrup, accomplishes the connection of the middle ear with the inner ear. As shown in Figure 2.2, the footplate of the stirrup attaches to the cochlea through a structure, called oval window.

The middle ear locates in an air-filled chamber. It is connected to the nose cavity through the Eustachian tube, which is $3.5cm$ long. Since the inner ear is filled with fluid, it is incompressible. Further the resistance of movement of fluid is much higher than air. The force provided by the vibration of the tympanic membrane needs to be mediated by the middle ear to maximize the

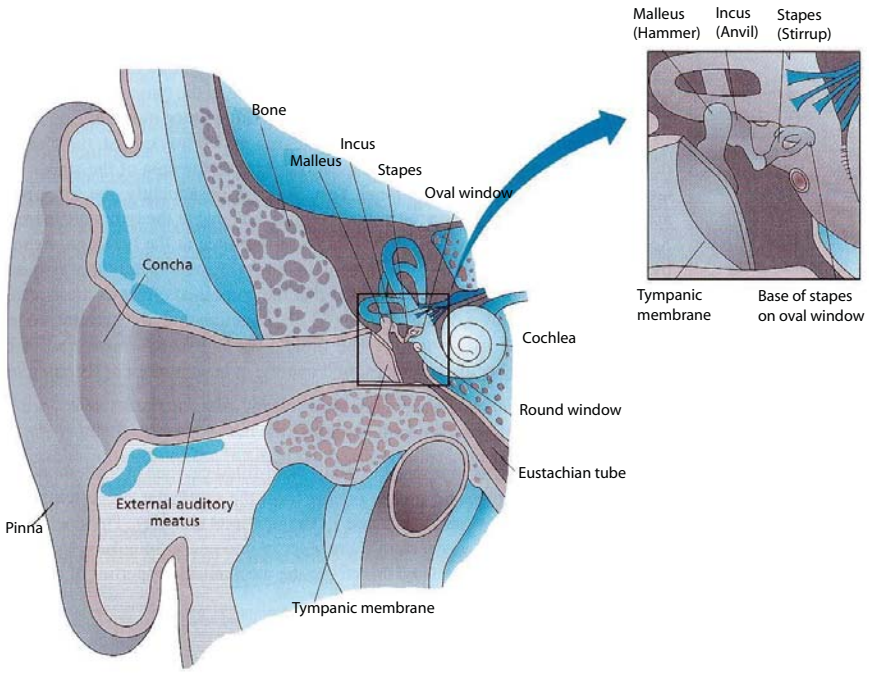

Figure 2.2: Components in human ear: outer ear; middle ear; and inner ear. Re-drawn from Purves et. al. (2001), based on [53].

transmission: firstly, these three bones form a lever function, which gives a lever ratio about 1.3; secondly the area ratio between the tympanic membrane and the area of stapes in contact with the oval window, is about 17. All in all the two effects increase the pressure with a ratio of 22, or 27 dB.

The Eustachian tube is normally closed, and its open will equalize the pressure inside the middle ear chamber and atmosphere. The blocking of this tube will stop the oxygen supply to the middle ear. When the remaining oxygen is absorbed by the tissue, the low-pressure takes place, which causes the inwards deformation of the tympanic membrane and the decreasing of hearing sensitivity. That is the main reason that passages should swallow or yawn while the airplane is landing, since these actions will allow the Eustachian tube to open and let the middle ear chamber has the same pressure as outside.

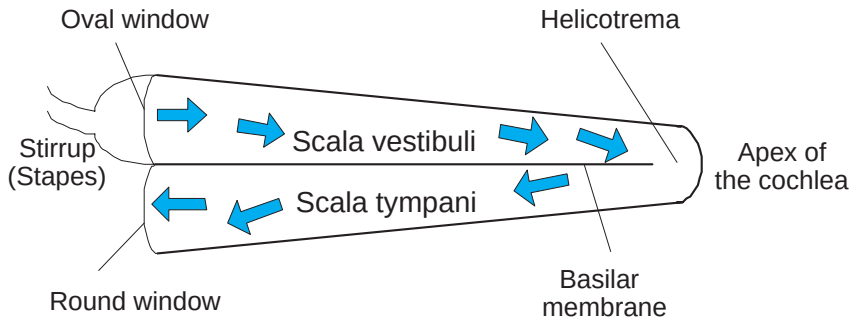

Figure 2.3: Simplified cross-section of an straighten cochlea. The inwards movement of the stirrup causes the flow of the fluid inside the cochlea, and arrows indicate the moving direction. Re-drawn from Purves et. al. (2001), based on [53].

2.2.3 The Inner Ear

The inner ear consists of a number of cavities in the temporal bone. Here we focus on the hearing sense organ: Cochlea. The cochlea is a snail-shell shaped structure, and is filled with lymph. It is only $10mm$ in diameter, however the total length from the base to top of the cochlea, if straightened out, is about $32-34mm$ long. There are two membrane-covered openings in the cochlea: the oval window and the round window. As just introduced, the oval window is the contact between the stirrup and the cochlea. The round window allows the fluid movement inside the cochlea and keeps the volume constant. Figure 2.3 shows the structure of a straightened cochlea tube [53].

The cochlea tube contains two chambers. They are separated by a thin bony shelf, called cochlear partition. The upper chamber, called scala vestibuli, leads from the oval window to the apex of the cochlear partition. The lower chamber, the scala tympani, extends from the apex of the cochlea to a membrane-covered opening, the round window. At the apex of the cochlea, the fluids in chambers can flow through a small opening: the helicotrema. The basilar membrane sits on the bony shelf of the cochlea. It begins with a narrow (about $0.1mm$ in width) and stiff form at the side close to the two windows, and ends with a wider and flexible form at the apex of the cochlea with the width of $0.5mm$. This membrane contains many thousands of stiff, elastic fibres, and it plays a big role in the hearing system.

A complex structure - the **organ of Corti** consists of the basilar membrane,

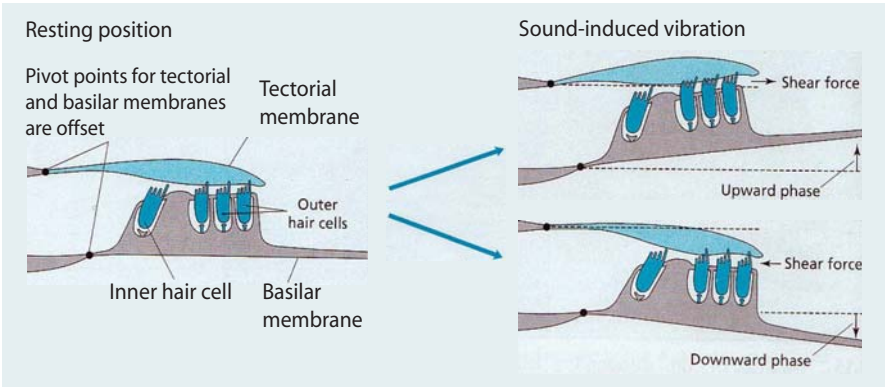

Figure 2.4: Basilar membrane displacement. Re-drawn from Purves et. al. (2001), based on [53].

the tectorial membrane, and hair cells, shown in Figure 2.4. Hair cells locate on the top of the basilar membrane, forming four rows or more along the length of the basilar membrane. Hair cells in the inner side of the cochlea spiral is called **inner hair cells**, and there are about 3500 hair cells. The rest is called **outer hair cells**, which are arranged in three rows in cat or up to five rows in human. In total there are about $12,000$ hair cells [60]. Hair cells are special nerve cells with small hairs protruding from the top of cells. A soft membrane, tectorial membrane, covers the top of hair cells. As shown in Figure 2.4, hairs of outer hair cells are embedded into the soft membrane.

Inner hair cells are the main sensory cells, and the sensory information of sound is conveyed by them. When the air pressure waves cause the vibration of tympanic membrane (eardrum), the fluid inside the inner ear starts to flow. This movement, as a consequence, causes the deformation of the basilar membrane. Hence, as shown in the right-hand side of Figure 2.4, the top of hair cells will bend back and forth. Since inner hair cells are connected to afferent fibres of the auditory nerve, the movement triggers the production of neural impulses.

Unlike inner hair cells, outer hair cells can expand, contract, and change their size. They are connected to efferent fibres, which convey signals from the central auditory system back to the cochlea. When the ear is exposed to weak sounds, the central auditory system will send signals to control the muscle tissue in outer hair cells, in turn will amplify vibrations of the basilar membrane so that the movement is big enough to stimulate the reaction of inner hair cells.

2.2.4 Short Summary of Auditory System

With the general idea about structures and functions of the ear, let us summarize human auditory system with the mechanical response of the ear while exposing to air pressure waves.

Due to the selectivity property of sensory system, human auditory system responds only to a particular range of stimuli: sound pressure wave frequencies between 20 Hz and 16 kHz. Speech signals nearly fall in the range of 200 Hz to 8 kHz. When a sound wave travels to the ear, the tympanic membrane starts to vibrate. Through the ear canal the stimulus is amplified by the outer ear. The vibration of the eardrum pushes the hammer of the middle ear, and consequently move the stirrup at the same frequency as the sound wave. The footplate of the stirrup touches the cochlea with the membrane-covered opening: oval window. The back and forth movement of the stirrup forces the window to move, which drives the fluid (lymph) to flow, from the scala vestibuli chamber through the small opening (helicotrema) to the scala tympani. Since the liquid is incompressible, the same amount of movement on the oval window will be reflexed back to the round window. The liquid movement inside the cochlea transmits the pressure from the middle ear, in the meanwhile deforms the basilar membrane. The displacement between the basilar membrane and the soft tectorial membrane will provide the shearing motion, which will displace the protrude of both inner and outer hair cells. Furthermore, the displacement of inner hair cells will trigger neural impulses by conveying hearing information to afferent fibres of the auditory nerve. The ear is equipped with several muscles or muscle-like tissues to cope with either low sound pressure level (SPL) or high SPL. Outer hair cells link the gap between the basilar membrane and the tectorial membrane. The mechanical coupling amplifies the displacement of the basilar membrane when a weak sound is presenting, and therefore increase the inner hair cells' response. This function is usually called cochlear amplifier. On the other hand, high SPL sound has the tendency to destroy part of our hearing functions. There are two small muscles attached to the hammer and stirrup bones in the middle ear, controlled by signals sending from the brain. When our brains realize that we are exposed to a SPL higher than 70 dB, signals will be sent out to stir these muscles, in order to descent movements of three bones in the middle ear. Unfortunately it takes about 25-150 $msec$ before the action takes place, therefore it can not protect the ear from impulsive sounds. Moveover high SPL sound can also destroy the cochlear amplifier, and the hearing loss is unreconstructed. Besides noise, the other unavoidable cause is aging.

The maximum displacement of the basilar membrane has been proven to be frequency-dependent by Georg von Bekesy in the 1950s. The special structure

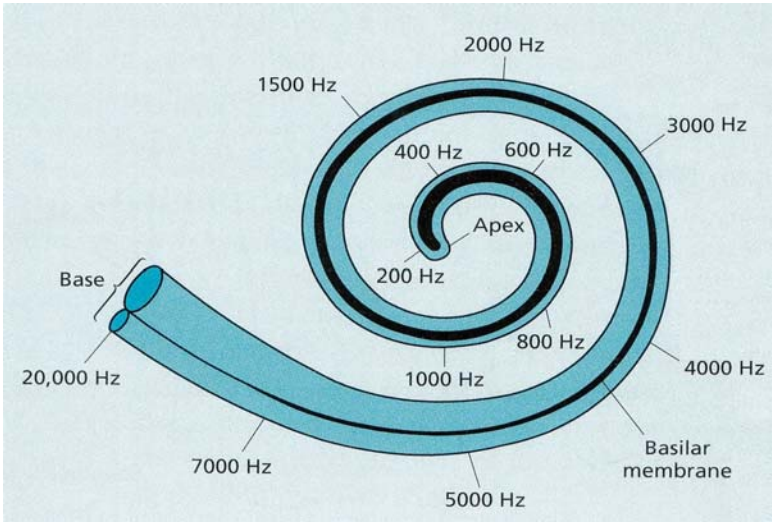

Figure 2.5: The frequency map of the maximum displacements along the basilar membrane. This figure is taken from FIG. 4.14 in [53].

of the basilar membrane introduced earlier, allows it to respond frequency at different place along the membrane: the low frequency triggers the maximum displacement near the apex; and the high frequency arouses the maximum displacement close to the two windows, as shown in Figure 2.5. In other words, a particular frequency will only give the largest deformation at a particular place of the membrane, and inner/outer hair cells at that location will be activated to trigger neural impulses.

Chapter 3

Overview of Cognitive Component Analysis

This chapter will provide an overview of the cognitive component analysis (COCA). The outstanding achievements of the machine and computer analysis on many perceptive tasks are our main motivations to propose COCA. Consequently, our hypothesis is composed, and it aims at explaining the relation between the human cognition and statistical regularities. Chapter 2 went through the functionality of human auditory perception and cognition. With this background in mind, we will construct the classic preprocessing pipeline of COCA for speech signal processing, and each step in the preprocessing will be explained in detail, which attempt to emulate some functionalities of human auditory system and the structure building in the brain.

3.1 The Hypothesis of COCA

The hypothesis of COCA is primarily based on two concepts: statistical independence and sparse representations. The independence can dramatically reduce the perception-to-action mapping, and a sparse representation has been found in representing sensory inputs, and is energy efficient. They will be introduced and explained in depth, which leads to our ecological hypothesis of COCA.

3.1.1 The Statistical Independence

The human cognitive system is able to model complex multi-agent scenery, and use a broad spectrum of cues to analyze perceptual inputs, so as to infer the proper action for a given situation. It is believed that an evolutionary optimized brain is capable of exploiting robust statistical regularities while making inference of appropriate actions [7]. In the article, Barlow has posed the question: *What is the source of the extensive and well-organized knowledge of the environment implied by the possession of an cognitive map or working model?* He hence argued that the sensory information received by the brain is not a totally correct answer, but rather statistical regularities in these messages must be recorded by the brain, in order to inform the brain what usually happens. Furthermore, he stated that an unsupervised learning algorithm can provide us with a factorial code of independent visual features; and our visual feature detectors are the result of reduction process on the redundancy of sensory messages, and these detectors are statistically independent. In [69], it has also been shown that the unsupervised learning can discover regularities in the input. The property of unsupervised learning will be discussed in detail in Chapter 4.

Barlow was thus led to propose that our visual cortical feature detectors might be the end result of a redundancy reduction process, in which the activation of each feature detector is supposed to be as statistically independent from the others as possible. Such a factorial code potentially involves dependencies of all orders, but most studies have used only the second order statistics required for decorrelating outputs of a set of feature detectors. The knowledge on an independence rule will allow the system to take advantage of the corresponding factorial code, typically of (much) lower complexity than the one pertinent to the full joint distribution. The exploration of the independency in the relevant natural ensemble statistics, has been carried out for more than a decade. Bell and Sejnowski have extracted 'independent components' from an ensemble of natural scenes, and proved the detectability of natural images' edges by linear filters. They anticipated the predictive power of abstract unsupervised learning techniques [8]. More studies of independence in primary sensory systems includes [35] on visual feature extraction from images, and [49] on natural sound coding, where sounds are categorized into three distinct classes: non-harmonic environmental sounds; harmonic animal sounds; and speech having both harmonic vowels and non-harmonic consonants. The representations found in human and animal perceptual systems, closely resemble the theoretically optimal representations from the unsupervised learning of perceived signals separation, namely independent component analysis (ICA).

ICA has been first brought to bear in [12]. Due to its generality, ICA has been used in many areas, such as textual information analysis, sound signal separa-

tion, image and biomedical data processing, etc. More derivatives have later been proposed, and the original linear ICA has been further modified to fit different applications. Such as non-linear ICA [37] for non-linear source separation; convolutive ICA [20], shifted ICA [61], and independent vector analysis (IVA) [47] to solve permutation ambiguities of sources. One of the classic applications of ICA model in signal processing is blind source separation (BSS). Speaking of BSS, people may think of the cocktail party problem (CPP), see e.g. [33]. CPP is to separate sound sources of different speakers and/or music, using recordings of one or more microphones, which scenario is often happened in a cocktail party. In the meanwhile of discussing the trade-off between timing and frequency analysis on sounds [68], the property of ICA in accounting for the neural response has been referred briefly. It claimed that for auditory system, ICA is invoked to elucidate the neural response properties at the very earliest stage of analysis; whereas ICA accounts for the response properties of cortical neurons in the visual system, which is the second stage after photoreceptors (visual information is received by photoreceptors, and transmitted by sensory pathway to the brain). What decides the stage of ICA analysis of sensory signals? One explanation states that ICA is applied at the point of expansion in the representation, for details see [68].

Besides the significance of statistical independence shown by the theoretical and computational optimal outcomes of ICA on perceptive tasks, let us seek some evidences of independence from biology and philosophy. Wagensberg has pointed out that the success of abstract 'life forms' is linked to their ability to recognize independence between a predictable and an un-predictable process in a given niche [83]. He claimed that some objects display certain rare property, which can be seen as the perpetuation of the objects, and he named these objects *living individuals*:

> *A living individual is a part of the world with some identity that tends to become independent of the uncertainty of the rest of the world.*

This represents a precision of the classical Darwinian paradigm by arguing that natural selection simply favors innovations, which increase the independence of the agent and un-predictable processes. The agent can be an individual or a group, and a group is seen as an association or a society composed by similar individuals in a broad sense. Living individuals fit themselves into hierarchical levels of the organization of living matter. It implies that in order to create a society or alliance, living individuals have to give up their independence for the benefit of a group, which in turns can increase the independence of the group as an entity.

The individual sacrifices some of its individual independence in exchange for belonging to a whole independence of which is compatible with the environmental uncertainty.

Due to the versatility of the statistical independence, it becomes the main basis of our hypothesis.

3.1.2 The sparse representation

Sparseness, like independence, has the property of reducing computational complexity. In sensory coding, 'sparse distributed' coding was invoked and proved to be near optimal in representing natural scenes in the visual system [24]. The studies have shown that the sensory information is encoded by a small number of neurons at a certain point of time. Field has argued for the importance of sparseness in which the above mentioned statistical independent feature detector is activated as rarely as possible. This coincides with Barlow's 'Minimum Entropy coding'.

The principal of sparse coding has a history of more than three decades. It has been suggested and studied from different viewpoints and reasons. A current review on the sparse coding [67] has summarized its advantages as:

- Sparse representations are the most effective means for storing patterns in the associative memories;
- Sparse coding clears structures in natural inputs;
- Sparse representations have advantages to designate complex data in a explicit and easy-to-read way;
- Spareness saves the energy required for signaling in cortical neurons w.r.t. the low average firing rates.

The theoretical and experimental studies of the sensory input encoding in cerebral cortex, have been carried out systematically. Especially, a series of study in visual system, e.g. [24, 8], has shown that the spatial receptive field properties match sparse representations, and by maximizing statistical independence and sparseness of the representations, the resulting receptive field properties share similarities with those of cortical neurons [66].

Does the representation of auditory perception of natural sound abide by the independent and sparse rules as well? Recent studies furnished some evidences

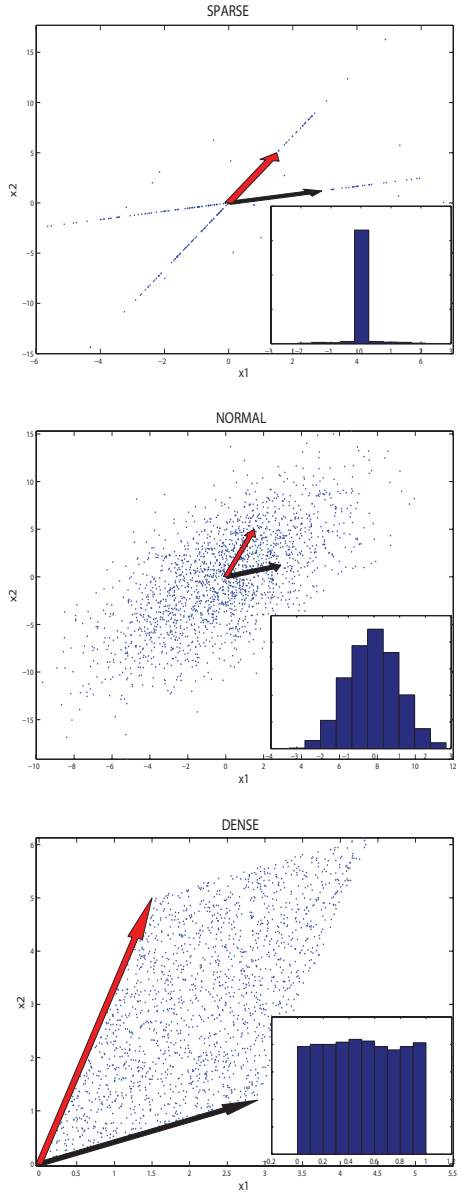

Figure 3.1: Prototypical feature distributions produced by a linear mixture based on two sources with sparse, normal, or dense histograms respectively. The characteristics of a sparse signal produces a 'ray structure', in which the ray is defined by the vector of linear mixing coefficients: One for each sparse source.

leading to an positive answer: The receptive field properties of auditory nerve cells invoke a strategy of sparse independent manner to represent natural sounds [49, 68].

The histogram of a signal can be coarsely described as sparse, normal, or dense, seen Figure 3.1. The upper panel shows a typical appearance of a sparse source mixture. The sparse signals are made of a few samples with relatively very large magnitude in a background of a mass number of small or weak signals (see the inlet in the upper panel). When mixing such independent sparse signals in a simple linear manner, we will most likely end up with a 'ray structure', which we consider emblematic for our COCA analysis. If a signal representation exists with a 'ray structure', ICA can be used to recover both line directions (defined by column vectors of the mixing matrix) and original independent source signals.

3.1.3 The Hypothesis

ICA as one of the growing statistical machine learning techniques has demonstrated the crucial importance of the *statistical independence* on a number of perceptive tasks. Unsupervised grouping of data by ICA, has been pursued earlier for abstract data including text, dynamic text (chat), images, and combinations [31, 32, 41, 42, 44]. Furthermore, as stated earlier, optimized representations of the low level cognition are known to be based on independence in the relevant natural ensemble statistics.

These findings evoke our query that *whether ICA is also employed by human brain in higher level cognitive activities*. During the process of seeking for an answer, the independent cognitive component hypothesis emerges, and it is based on the statement that *the characteristics of human cognition are determined by statistical properties of the input*, and runs: *human cognition uses information theoretically optimal ICA representations for generic data analysis*. Based upon evidences of independence and sparseness from a series of theoretical, computational and experimental research, especially in the psychology and physiology standpoints, our hypothesis is ecological: we assume that features that are essentially independent in a context defined ensemble can be efficiently coded using a **sparse independent** component representation. COgnitive Component Analysis (COCA), as a straightforward outcome, is proposed, which is defined as the process of unsupervised grouping of generic data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity, see [30].

The way we represent the results of human cognitive activity, is intuitive. Since the mechanisms of human cognitive activity are still not fully understood, to

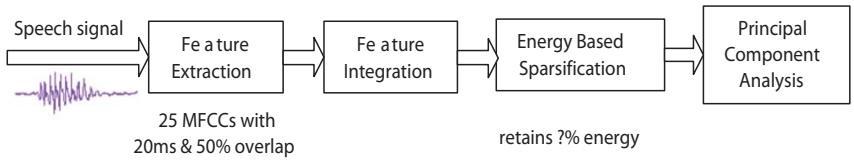

Figure 3.2: Preprocessing pipeline for COCA of speech. Feature extraction is normally followed by feature integration, so as to obtain features at longer time scales. Energy based sparsification aims at reducing the intrinsic noise and getting sparse representations. PCA projects features onto a base of cognitive processes. A subsequent ICA can identify the actual ray coordinates and source signals.

quantify cognition may seem ambiguous and may also be considered way too ambitious. However human behavior, as the direct consequence of cognition, contains rich phenomenology, and is easier to access and model than human cognition. As to speech relevant cognitive activities, we pay close attention to the human behavior on sound (speech) perception, speech context understanding, and judgment based on speech. Hence we represent human cognition simply by a classification rule, i.e. based on a set of manually obtained labels we train a classifier using supervised learning. Manually obtained labels reflect human judgment of a given tasks. *The question is then reduced to looking for similarities between representations in supervised learning (of human labels) and unsupervised learning that simply explores statistical properties of the domain.* If the representations provided by both unsupervised learning and supervised learning methods coincide with each other with high correlations at different levels, we can conclude that the evidence show positive support on our hypothesis about the consistency of statistical regularities/independence (unsupervised learning) and human cognitive processes (supervised learning).

3.2 Preprocessing Pipeline of COCA

For any system, data preparation is the first step. Usually directly working on raw data is not optimal, and representative information for various tasks needs to be extracted in a efficient way, in order to prepare informative data for the system. This step is normally called feature extraction. To emulate how brain processes speech signals in the early stages, we design the preprocessing pipeline, starting from feature extraction, shown in Figure 3.2.

Hence, in this section we will elaborate on the steps in the preprocessing pipeline. The design of the preprocessing procedures focuses on two aspects: on one hand, the flow-chart is built in respect to standard techniques using in speech signal processing; on the other hand, the choice of techniques are in connection with the human auditory system response.

3.2.1 Features

To choose features which are capable of representing the information that human ear perceives, we need to briefly delve into the physiology of human ear, psychophysics and psychoacoustics.

Does the human ear work as a Fourier analyzer?

Sounds are detected by ears as frequencies. As briefly introduced in Section 2.2, air pressure waves are transformed as the vibration of tympanic membrane at the same frequency as the waves, and these frequencies are converted to the displacement at corresponding places in the cochlea (basilar membrane). The displacement causes inner hair cells at the certain area to bend, which in turn will trigger auditory nerve fibers to produce neural impulses. The sound frequency that produces the largest response from a particular nerve fibre is called *characteristic frequency*. A nerve fibre will response to a broad range of frequencies, however the maximum response happens only if the sound frequency matches this nerve fibre's *characteristic frequency*. The phase and intensity of a sound wave are claimed to be reflected by the pattern of firing in auditory nerve fibres. Therefore some researchers claim that loosely speaking, the ear is behaving like a Fourier analyzer, where each sound can be decomposed to a collection of sine frequency components [53].

To some extent, humans seem to be able to perceive the harmonics individually from a periodic sound, according to *Ohm's acoustical law*. Even though it does not apply when a complex tone is presented, in this case what we hear can be just a single pitch, we can indeed hear two separate tones while two simultaneous pure tones with differing frequencies are perceived. Hence we perceive sounds in terms of their Fourier components [60].

Non-linear frequency perception

Furthermore, humans do not perceive pitch or frequency of a tone in a linear manner. How do we perceive frequencies of sounds? The study brought the definition of 'mel' [76]. It was measured in a similar way as we measure human sensory magnitude. The subject was initially presented with a reference

stimulus, i.e. physical frequency 1 kHz, and referred it as 1000 mels. Other frequencies were presented at a time afterwards, and the subject needed to estimate the stimuli with regards to the reference stimulus, and assign a number. If the stimulus is perceived as twice as the reference, then it is labeled as 2000 mels; if perceived as half the reference, then 500 mels, and so on so forth. The experimental result shows that mel-scale is near linear below 1 kHz, and logarithmic above.

Critical band

We will dig into the cause of the following phenomena: in some cases noise dramatically degrades the hearing intelligence, in other cases noise has less affect on our signal perception. Here noise refers to a sound containing a broad range of frequencies with random phases but equal amplitudes. The explanation will guide us to understand the principle of auditory perception in the psychophysical way.

From the psychophysical studies of frequency masking, one opinion declares that humans use a function like a band-pass filter to perceive signals, select frequencies within the bandwidth, and remove the rest. If this opinion is accurate, we can explain the above mentioned phonomania w.r.t. noise masking. A sound signal will be effectively detected by a bandpass filter centered in the same frequency as the signal. For the same reason, if the noise locates in the pass band of the filter, the noise will affect our signal detecting ability. The assumptions for noise masking can be summarized as follow: first the presence of a signal actives certain auditory filters at its frequency; secondly when this activation excesses a threshold, the signal is successfully detected; thirdly the noise mask also actives some auditory filters. Consequently, if the activated filters by the signal are not the same ones activated by noise mask, the noise has no affect on signal detection; otherwise if the same filters are activated by both the signal and the noise, the detectablity will be impaired. By increasing the noise bandwidth, the signal detectability will become gradually decreasing. Nevertheless when a certain threshold of bandwidth is reached, the detectability will stay constant. Studies following this line support a view that the function of human auditory system for signal perception is fulfilled by a bank of bandpass filters from low frequencies (e.g. 20 Hz) to high frequencies (e.g. 16 kHz).

An explanation from the physiological point of view is focused on the basilar membrane. As introduced the *characteristic frequency* of nerve fibers linked with inner hair cells, each point on the basilar membrane thus can be regarded as a bandpass filter with a center frequency corresponding to the *characteristic frequency*, and a bandwidth [60].

Feature Extraction

Feature extraction is usually the first stage and the most important stage in a system, and it is influential to the overall performance of the whole system. Features are basically extracted from short time scales. The frame size may depend on applications, in other words features at different time scales may contain different information. A small frame size may result a noisy estimation; on the contrary a long frame size may lose the appropriate information in need. Later we will discuss the rule of the time scale.

To represent speech signals for machine speech analysis, spectral features of fairly low dimensionality are usually used. These $20 - 30$ dimensions of features are usually uncorrelated. The basic features in COCA analysis are extracted from a digital speech signal leading to a fundamental representation that shares similarities with human auditory system. These so-called mel-frequency cepstral coefficients (MFCCs) are well-known in speech and speaker recognition society. MFCCs are designed as perceptually weighted cepstral coefficients, since the mel-frequency warping emulates human sound perception. They have been developed for speech processing, e.g. speech recognition and speaker recognition [72]. However MFCCs recently have been popular in many other areas, such as music genre classification [2], audio similarity measure [6] and instrument classification [63], etc.

Due to the non-stationary property of speech, the basic features need to be extracted from audio signals in, e.g. $10 \sim 40$ $msec$, in which period the signal is assumed stationary. Here we name these short-time features the basic features in COCA analysis. The computation of MFCCs is based on the short-time time-frequency analysis. MFCCs decompose signals into broad spectral channels, and compress the loudness of the signals. The block diagram for computing MFCCs is given in Figure 3.3. Other methods to implement MFCCs exist, and different implementations have been compared in [87]. The fast Fourier transform (FFT) transforms the convolution relationship between excitation sequence and the vocal system impulse response into production in the frequency domain; and the logarithm, afterwards, provides us the linear combination (addition between these two). The mel-frequency warping changes the frequency scale from linear to mel-scale, which attempts to mimicry non-linear human pitch perception. The mel-frequency warping is realized by a bank of bandpass filters, termed critical band filters. A few types of filters can be used, such as triangular shaped filters, hanning filters and hamming filters. Here triangular shaped filters are in use, and center frequencies spacing of the filters follows Mel scale. Loosely speaking, critical band filters represent the frequency resolution of the peripheral human auditory system, and they also reflect the auditory system in a way that signals passing through different critical bands are processed independently [29]. Finally discrete cosine transform (DCT) brings us to the mel-cepstrum. For detailed description, see [16].

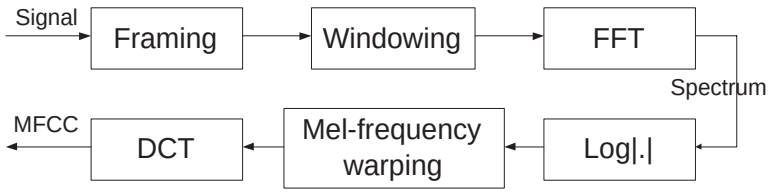

Figure 3.3: Block diagram of MFCC

MFCCs follows the Fourier transform and mel-frequency scale. All in all, MFCCs share two aspects with human auditory system: A logarithmic dependence on signal power and a simple bandwidth-to-center frequency scaling so that the frequency resolution is better at lower frequencies. Therefore they can loosely represent the human auditory response, except for part of the outer ear, which is critical for sound localization and loudness accuracy.

3.2.2 Feature Integration

In multimedia analysis, for a feature to reveal the semantic meaning of an audio signal, analysis over a much longer period is necessary, usually from one second to several tens seconds [84]. Feature integration is by and large the way to combine the information from several short-time features (vectors) into a long-time feature (vector), in order to capture the representative information for a certain task. In [84] short-time features were named *frame-level features*, and long-time features the *clip-level features*. The clip-level features can characterize how frame-level features change over a clip. Here we will introduce the current existing feature integration methods in audio signal processing area. It is by no means a comprehensive review, but rather to give a general access.

Review of Feature Integration

Feature integration is referred to the process of constructing features at a longer time scale than the basic ones, so as to obtain discriminative information for a given task. Constructing a feature at a longer time scale often involves sample compression, since a single sample at longer time scale is extracted from several samples at a short time scale.

During the course of searching for appropriate features for various classification or recognition tasks, researchers have realized that the systems performance is hard to be further improved by only using short-time features. The idea

of introducing features at longer time scale is relatively new. As one of the first articles discussing the integrated features in audio analysis, Wold et al. [85] among other work in the relevant area has brought the upsurge of feature integration. Instead of referring directly to feature integration, clip-level features have been divided into: volume based, ZCR based, pitch based and frequency based [84]. More often the mean and variance of several short-time features are used as the clip-level features [81]. Zhang and Zhou [86] have focused on decreasing the false alarm rate of audio segmentation. To enhance the system performance, a rough segmentation step based on large-scale classification was introduced to the system before a subtle segmentation, and the chosen features contained several large-scale features (e.g. low short-time energy ratio, high zero-crossing rate ratio, and harmonious degree) and the mean and variance of short-time features. A frequency band approach of feature integration has been proposed in [55]: the basic features were calculated from 23 $msec$ half-overlapping frames of audio signals. A power spectrum was then calculated crossing 64 consecutive frames on each basic feature dimension individually, and the resulting time scale for the integrated features was 743 $msec$. Finally the energy of features was summarized into 4 frequency bands. Along with feature integration methods, two new feature sets were also introduced: psychoacoustic features (roughness, loudness and sharpness) and auditory filterbank temporal envelopes. These features were developed to throw light upon different aspects of human auditory system.

A few statistical models have been applied to feature integration. Multivariate Gaussian model, mixture of Gaussians (MoG) can be such models. The multivariate autoregressive model (MAR) has been recently used in music genre classification [57]. Compared with methods using Multivariate Gaussian model and MoG, MAR is able to capture both the temporal dynamics and the dependencies among the short-time feature dimensions.

Feature Stacking

Among all feature integration methods, a simplest approach is to stack short-time features into a long vector. Feature stacking has been used in audio retrieval and indexing to obtain long-term spectral characteristics of short-time MFCCs [75]. Stacking can be expressed as below:

$$\mathbf{v} = f(\mathbf{M}), \tag{3.1}$$

where the function $f(\cdot)$ carries out the stacking operation, in other words vector concatenation. It stacks all column vectors in matrix $\mathbf{M}$ $(d-by-n)$ into a single column long vector, and the number of column n in matrix $\mathbf{M}$ decides the time scale of the new feature vector $\mathbf{v}$. The dimensionality of $\mathbf{v}$ will be $d * n$. In the later stages of the preprocessing pipeline, dimensionality reduction algorithm

will be used to select the most important dimensions. Figure 3.4 illustrates the stacking procedure used in COCA analysis.

1. Firstly, digital speech signals are truncated into short-time frames. Here the basic features are based on short-time speech signals, e.g. 20 $msec$ which corresponds to 320 samples at 16 kHz sampling frequency. A certain overlap between two adjacent frames is set;
2. Since the side lobes of the rectangular window spectrum cause signal power to 'leak' into other frequencies, a hamming window is applied to each frame;
3. d-dimensional MFCC is extracted from each frame, which forms a d-dimensional vector. e.g. a 25-dimensional short-time feature vector;
4. According to the new time scale, the matrices $\mathbf{M}$ are formed by the n MFCCs starting from the first n frames, and the frame overlapping among matrices is optional. The residual which is not enough to form a n column matrix at the end of the signal, is excluded;
5. Each matrix is stacked following Equation 3.1 into one $d * n$-dimensional vector.

The $d * n$-dimensional features extracted with 50% overlap among short-time frames and no overlap among matrices, represent speech information at a long time scale of $20msec * (n + 1)/2$.

3.2.3 Energy Based Sparsification

As mentioned in Section 3.1 the receptive field properties of auditory nerve cells invoke a strategy of sparse independent manner to represent natural sounds. Hence we here carry out the energy based sparsification (EBS), and it is also meant to emulate the cognitive process: 'attention', in a way that strong (loud) signals win awareness.

> *"Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others, and is a condition which has a real opposite in the confused, dazed, scatterbrained state which in French is called distraction, and Zerstreutheit in German."* [38].

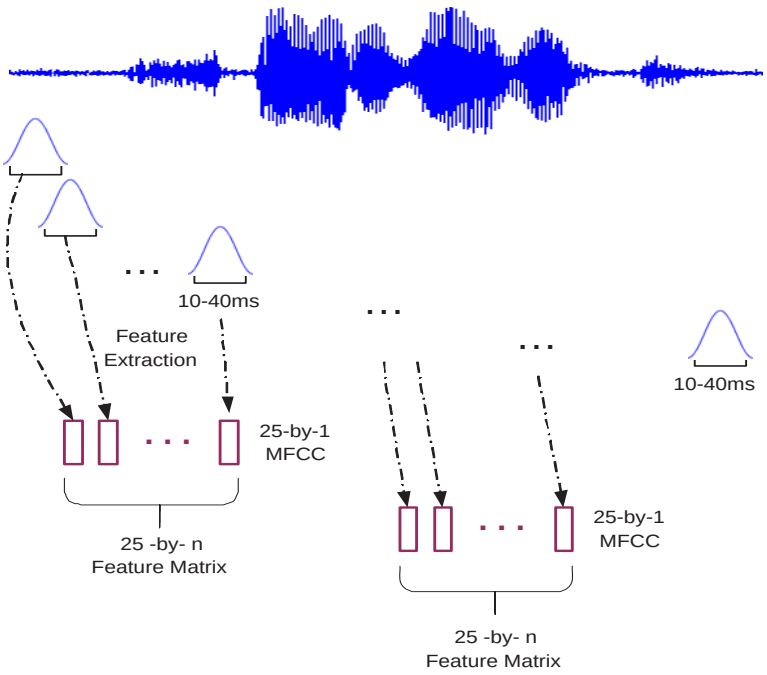

Figure 3.4: Speech feature extraction and stacking.

Attention is the ability of concentrating on one event in the surrounding while ignoring other events. Concentrating on one person speaking or one conversation in a noisy environment is one example. Cocktail party problem involves both attention concentration and attention shift, e.g. you shift your attention while somebody outside your conversation is calling your name. Here we only borrow a limited interpretation of the concept: 'attention', and represent the cause of 'catching attention' as the form of signals with large magnitude (energy) in the background of many weak signals.

EBS is a simple way to filter out weak signals, and it emulates the **detectability** and **sensory magnitude** from perceptual principles [53]. Detectability in perceptual principles means the ability of sensory organs to detect the environmental stimulus. It depends highly on the intensity of the stimulus and the variability of neural signals as well. Figure 3.5 shows a picture of a typical neuron. Neurons have many shapes and sizes, but they share some common structures. Dendrites are the entrance of inputs signals, and their branching structures connecting with outputs of other neurons, allow them to receive signals from a number of different synapses. Nucleus sits inside the cell body, and controls the functions of neuron together with other organelles. Axon is the signal transmitting channel. Axon ends with a mass number of terminal endings, which usually connect to dendrites of other neurons or directly to muscles. Terminal endings work as the output exit. Actually dendrites and terminal endings are not physically connected, and gaps between them are crossed by chemical signals, called neurotransmitters. When a stimulus is sensed, neurotransmitters are generated. If they are larger than a certain threshold of the dendrites, the cell will fire, and consequently will send signal through its axon to the terminal endings. In turn neurotransmitters will be released to the dendrites of other neurons [71]. Signals from a sensory organ, need to travel through a series of synapses from low level to higher level of neural processing, until reaching the corresponding cortex. For hearing system, there are five synapses starting from the hair cells to the auditory cortex. These stages do not only transmit signals, but also transform them into a refined form, such as selectively retaining useful information and discarding less important information. All in all the relationship between the intensity of a stimulus and the dendrites threshold will influence the detectability. Therefore sparsification is done by thresholding the stacked features. Since MFCC coefficients are energy based, the thresholding is applied directly on the amplitude of the coefficients, and only coefficients with superior energy than the threshold are retained, and the rest is set zero. In our study, thresholds are set empirically.

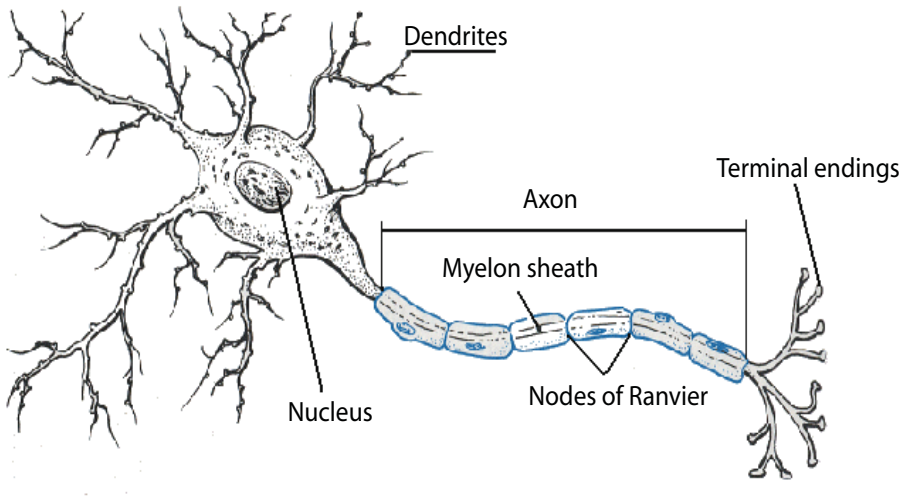

Figure 3.5: A typical Neuron.

3.2.4 Principal Component Analysis

COCA is a generalization of principal component analysis (PCA) based 'latent semantic analysis', originally developed for information retrieval on text [15]. PCA is an orthogonal linear transformation technique. It is often used for dimensionality reduction, which transforms and projects the data to a new coordinate system with lower dimensions, and in the meanwhile remains the most variance of the data.

In textual information analysis, latent semantic indexing (LSI) or latent semantic analysis (LSA) assumes that semantic content of the text, e.g. a paragraph, even a whole document, can be reflected by the sum of the meaning of words it includes. This assumption successfully avoids the complex syntactic problems, and converts the semantic indexing to a corpus based problem. The same word can have distinguished meanings in different context or used by different users, and this kind of words are called polysemy; and different words may also mean the same depending on the context, and they are called synonymy. The latter shows the variability of expression to refer to the same object. It is highly dependent on the context, users knowledge, and linguistic habits, the so-called idiolect. It has shown that the possibility for two users to choose the same word for describing a single well-known object is less than 20% [26]. The polysemy indicates the various referential significance of one word. For information re-

trieval, the task is to match the words of queries with words of documents or the conceptual content of documents. Since the words in a search query are not always included in the aiming documents, or in some cases the words of query may be covered in some irrelevant documents (where different meanings of the words have been refereed to), to discover the latent semantics is indispensable.

Normally, text data are formed as a large term-document matrix. Terms are the representative words in the documents. The matrix can be decomposed by some statistical machine learning techniques, and also can be projected into a low-dimensional 'semantic' space from the original high-dimensional space. In this low-dimensional space, words are seen as points, and meaning is represented as vectors [74, 15]. Therefore the position in the space is served as indexing, and documents having similar or common topics locate close to each other in the space. Since LSA tends to utilize the semantic meaning, rather than the word appearance, two different words can sit very close in the space representing similar meanings. The similarity between documents in the space is usually measured by their cosine. A new document or paragraph can be represented by a new vector of the words, and the position reveals the semantics of the text. LSA successfully solve the synonymy problem, however the polysemy problem may seem a bit harder to be fully solved. A word in the semantic space is a single point. For polysemy the weighted average of all the meanings it may have, will decide the position of the point in the space. If a certain meaning of the polysemy is far from the averaged meaning, LSA will find it hard to indicate the referred meaning of this word in current usage. A example of text analysis will be given in Section 3.3, and more detailed processing of text data will be presented.

The resulting low-dimensional space is regarded as the basis for all cognitive processing [40]. Some cognitive scientists believe that the performance of LSA resembles humans performance in the way meaning is represented. Since LSA has human-like performance in text analysis, we envision that it can as well be used to get the relevant basis for cognitive related tasks, e.g. speech perception. It has been proved that in some cases, LSA can provide good simulations of human cognitive processes alone, and in other cases it is often operated as base for cognitive processes. Here we adopt this well-understood concept, PCA/LSA, as the knowledge basis of COCA analysis, and use other ways to transform it.

To grasp the essential information and discard the redundancy (e.g. noise) in the data, singular value decomposition (SVD) is invoked to select the most informative and important dimensions in a sense that maximal amount of variance is retained. The mathematical express of SVD on the data matrix X is

$$\mathbf{X} = \mathbf{U}\Lambda\mathbf{V}^T, \tag{3.2}$$

where $\mathbf{X}$ is a m-by-n matrix; $\mathbf{U}$ is a m-by-m orthonormal matrix; Λ is a m-by-n matrix with singular values along the diagonal; and $\mathbf{V}$ is a n-by-n orthonormal matrix. The dimensionality of data is reduced by projecting the data to the first k principal components ($k < m$):

$$\mathbf{Y} = \mathbf{U}_k^T \mathbf{X} = \Lambda_k \mathbf{V}^T. \tag{3.3}$$

SVD minimizes the distance between the projected matrix $\mathbf{Y}$ and the original matrix $\mathbf{X}$:

$$\|\mathbf{X} - \mathbf{Y}\|_2, \tag{3.4}$$

where the 2-norm of matrices is equivalent to Euclidean distance of vectors.

For visualization, the most important dimensions depend on applications, and are not necessarily the ones providing the greatest amount of variance in the original space, but the ones providing the best retrieval effectiveness. A few examples will illustrate this phenomena in the next section.

3.3 Where Have Cognitive Components Been Found?

Before we touch upon the research findings achieved by COCA of speech, let us go quickly through the research results by applying COCA on several topics: text analysis, music genre and social networks, to reveal the corresponding cognitive components. This section is based on [30] and appendix D, and aims to illustrate the generality of COCA.

3.3.1 Text Analysis

The vector space representation proposed by Salton [74] has promoted the development of statistical modeling for text analysis. A term set needs to be chosen from all the appearing words, and some common words, like 'am', 'is', 'are', etc. have to be excluded. Then a document is represented by a vector of term frequencies. Thus a term-document data matrix can be formed. Words are represented as data points in the vector space, and they can also be seen as vectors; and documents are seen as the combination of all the word vectors included in the document. Information retrieval and text classification in the vector space are based on the assumption that documents sharing similar topics locate 'close'

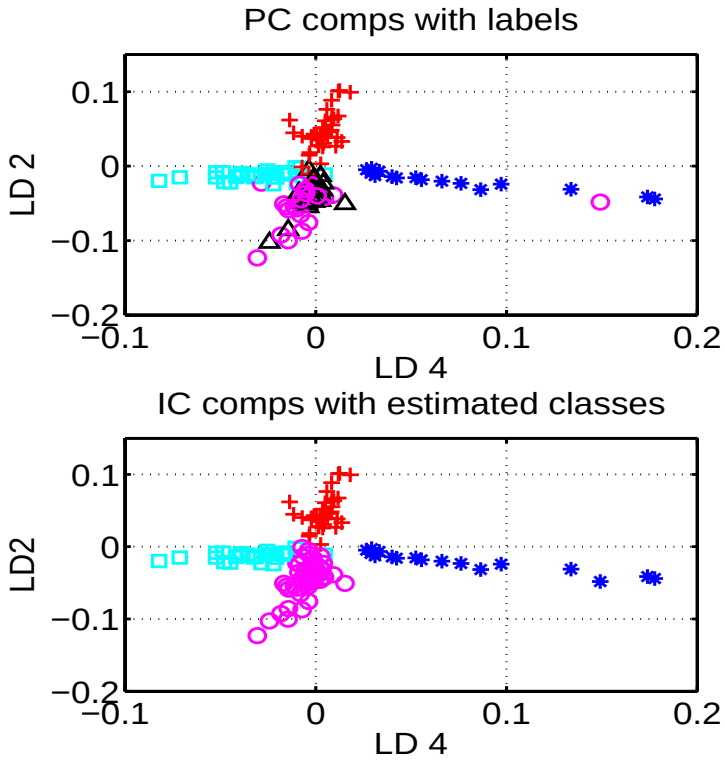

Figure 3.6: The scatter plot of text data set with five topics on two latent dimensions. Data are denoted with different shapes and colors indicating five independent components resulted from ICA algorithm on LSA coefficients. 'Rays' are mixed in a linear manner.

in the space, since they have similar semantic context in term usage. How to measure the semantic relatedness in the vector space? The cosine between vectors always comes in handy. The cosine value varies from -1 to $+1$: if the angle between two vectors is zero, cosine is 1 meaning identical; if the angle is 90^0 or 270^0, cosine is 0 meaning unrelated. However for properly normalized vectors, Euclidean distance is enough to explain the likeness. The length of the vector usually represents how much information it contains. Based on this idea, long text should have longer vectors than short text, and well-defined words have longer vectors than function words. An example is that word 'the' only has vector length of 0.03, however word 'horse' has 2.49 [40]. This also explains why the function and common words should be excluded in term selection. Normally the original high dimensional vector space of the data matrix is too noisy. To

focus on the essential semantic information in the corpus, the term-frequency vectors are projected to a lower dimensional space, determined by SVD of the data matrix. By this means we hope that the noise can be filtered out, and a stable core term set for each topic will be revealed. Wherefore the meaning of a new vector is defined by the interaction of other core term vectors close to it.

LSA restricts eigenvectors of a covariance matrix to an orthogonal basis, and it limits the interpretability of the representation. Due to this constraint, LSA is often used as a dimensionality reduction method, which is commonly followed by some post-processing techniques to relax the constraint. Figure 3.6 shows the scatter plot of a five-label text dataset in the latent semantic space. This data set covers five topics with large overlap of vocabulary. Since the term-document matrix is a sparse matrix, no sparsification has been implemented. The upper panel shows the projected data onto latent dimension 2 and 4. Data are tagged with true labels. Afterwards 4-component ICA was applied to reveal relevant cognitive components, shown in the lower panel. Data are denoted with estimated labels by ICA classifier. The 'ray-structure' is quite obvious, which is a signature of cognitive components. One independent components represents two topics, meaning they contain similar context. In LSA based ICA, topic vocabularies can have large overlaps. We envision that these implemented by overlapping receptive fields can detect more subtle differences than 'orthogonal' receptive fields.

3.3.2 Music Genre

With the rapid expansion of digitalized music on the internet, music information retrieval has become an important research topic. Due to the vast amount of music data, computational efficiency is a main concern. Music information retrieval is of interest both for commercial and academic reasons. Applications within content-based retrieval cover music instrument detection and separation [63]; automatic transcription of music [34]; melody detection [3]; musical genre classification [58]; sound source separation [82]; and singer recognition [80].

Here we are interested in music genre related cognitive component revealing. Musical genre classification is normally carried out by supervised learning on short-time features or integrated long-time features. It is a high level cognitive activity due to the ambiguity of the definition, and it also depends on people's background knowledge about music and individual opinions. For subtle classification of genres, it is rather subjective. Human performance (music experts not included) on genre classification is not so robust, and on one experiment it was around 52% accuracy among 11 genres, while the computer performance was about 44% [58]. A small set of music pieces is studied here with unsupervised

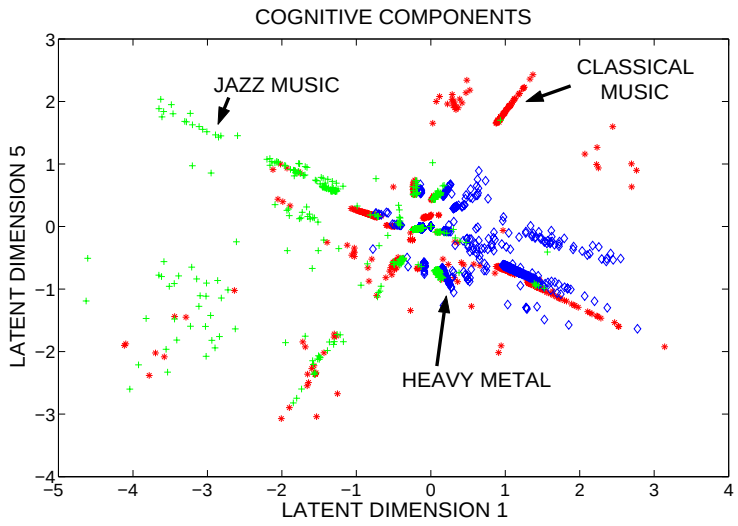

Figure 3.7: A latent semantic analysis like scatter plot of music data set with three genre labels: heavy metal, jazz, and classical music. Data are denoted based on the original manually given labels. The 'ray-structure' is striking, even though it is not a simple one-to-one correspondence to genres. We speculate that samples from longer time scale may help clear the ambiguity.

modeling, and this experiment aims at testing the possibility and limitations of unsupervised learning on high level cognition. A three-tune music set including heavy music, jazz and classical music, has been represented by spectral features, and each 13 dimensional feature vector is extracted from a music piece of 30 *msec* long, with 1/3 overlap. Since music reaches about 22 kHz, the sampling frequency is 44.1 kHz. In [2], authors claimed that MFCCs are relatively pitch independent. Based on the same principle introduced earlier in Section 3.1: sparse representations, MFCCs are sparsified. PCA projects the data into the latent semantic space. Figure 3.7 shows the scatter plot of data on first and fifth latent dimensions. Data are denoted in different shapes and colors according to the manually obtained labels. The 'ray-structure' has once again been revealed. As you may notice both in this illustration and the previous one on text analysis, the chosen dimensions for projection are not necessary the first ones providing the greatest amount of variance, but the ones providing the best retrieval effectiveness. Unlike text analysis, the genre representation is more complicated, and is not one-to-one correspondence. This representation is based on temporal scale of a frame, i.e. 30 *msec*. Further research on genre recognition has shown that integrating a number of basic frames into a feature

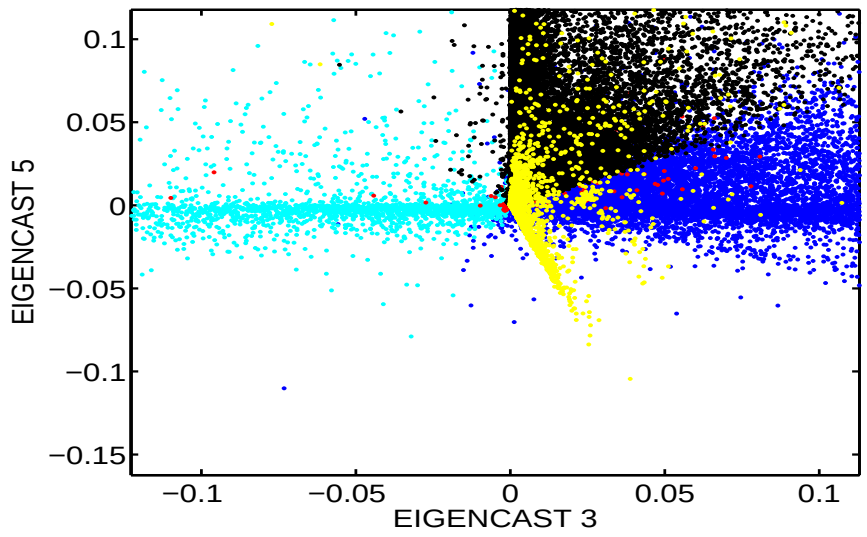

Figure 3.8: A LSA like scatter plot of social network matrix projected on two 'eigencasts'. Samples are coded according to 5 ICA components. A wider spreading 'ray-structure' can be seen. A simple unsupervised learning provoking independence can locate independent communities in complex networks.

vector representing longer time scales could enhance the classification performance [56, 58, 57]. This finding motivates our work on time scales in speech analysis, which will be covered later in Chapter 5.

3.3.3 Social Network

Part of the research in network science is about studying statistical properties of complex networks. Speaking of network, the world wide web (WWW) is often the first network, which rings the bell. The study of WWW [5, 4] has a little longer history than citation networks [48, 70]. Research has been focused on the properties that seem to be common to many networks: the small-world property, power-law degree distributions [1], and network transitivity. Here we attempt to study the property of community structure in a social network using unsupervised learning method: namely independent component analysis. Community in networks refers to a module, whose nodes are tightly knitted inside the module, and the connection between module are much looser.

The social network we are studying is actor to actor network associated with co-participation in movies. The data construction is similar to text analysis. Each movie is represented by an actor-list vector, and we call it the cast. Therefore actor can be seen the same as term in text analysis. Like term-document matrix, an actor-movie matrix can be constructed. Actors participating in a certain movie will score '1' as the entry to the matrix. The data cover 128.000 movies and 382.000 actors. By studying the actor-actor co-variance matrix provided by PCA analysis, we can find out the eigenvalues, called 'eigencasts' here, which reveals the community of actors who tend to co-operate in movies. The sparseness of the network decides that the most prominent variance components are related to near-disjunct sub-communities of actors having many common movies. Figure 3.8 gives the scatter plot of data projected on two 'eigencasts', and points are coded according to ICA components. At first glance, the structure is not so obvious. However a closer look at the origin area of the coordination system, reveals the linear mixture of sparse signals. The 'ray-structure' is still observable emanating from $(0, 0)$ with wider spread within each ICA component.

3.4 Summary

In this chapter, we gave a detailed introduction of cognitive component analysis, including the theoretical background, the hypothesis, and the motivation, etc. COCA is defined as the process of unsupervised grouping of generic data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity.

The COCA hypothesis was built upon statistical independence and sparseness, both have concrete theoretical evidences from physiology, biology, neuroscience and machine learning. Based on the knowledge of human perception, especially the auditory perception covered in Chapter 2, we attempted to emulate the human auditory system, and to process speech signals in a designated preprocessing pipeline: starting from feature extraction; feature stacking; energy based sparsification; to principal component analysis. Each step was built in order to approach human-like response. As the basis of the cognitive processing, PCA brings us to the knowledge basis of COCA analysis.

COCA has a broad generality. Section 3.3 illustrated its feasibility on various cognitive tasks. The ubiquity of 'ray-structure' representations was revealed and proven. We demonstrated that cognitive machinery developed for analyzing complex perceptual signals based on independency and sparseness, could be used to discover 'independent' document topics, to distinguish music genres, and to locate independent communities in complex networks. These activities

all involve higher brain functions. Furthermore, ICA has shown its capability of discovering the right representations.

Chapter 4

On Low-level Cognitive Component Analysis

The definition of cognitive component analysis along with its hypothesis has been elaborated in Chapter 3. The overall proposal of COCA and its foundation is associated with the mechanism of human cognition, and the scope of this dissertation: COCA of speech, spurs us to pay careful attention on human auditory system response to speech inputs. In the following two chapters, we will go along the development of COCA analysis of speech signals during the period of this dissertation, and open the course of COCA with low-level cognitive component analysis (appendix B, C and D), which gradually leads us to a more sophisticated completed analysis scheme covered by Chapter 5.

Section 3.3 has illustrated the ubiquity of 'ray-structure' representations in various cognitive tasks, and the statistic independence and sparse coding have successfully assisted COCA to reveal cognitive components of semantic context of text, music genres and co-working communities in actor social networks. All these illustrations are based on unsupervised learning scheme, which is counted on to show statistical regularities. Hence this chapter will focus on unsupervised grouping of data. First of all machine learning will be introduced, and more effort will be given, of course, to one of its sub-field: unsupervised learning. The description will be also from the point of view of probability theory and information theory. This layout will be followed by two applications of COCA

of digital speech signals, including 'fingerprint' of phonemes and 'voiceprint' of speakers.

4.1 Machine Learning

Machine learning is the research field concerned with the study of learning systems. It is devoted to the design and develop of techniques based on mathematics, statistics, engineering, computer science, and cognitive science, etc., to allow machines and computers to 'learn'. To infer human-like behaviors is the main theme of artificial intelligence. Automatic machine operation to select running modes in order to cope with different situations without human interaction is the ultimate goal. Machine learning in general is a generic term of numbers of learning techniques, and it can be categorized based on the data type, model structure, the purpose, usage, and so on. Here we will introduce machine learning following one taxonomy: supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning and game theory.

Supervised learning for a given task, builds a function based on training patterns. The training data usually come in pairs $\{\mathbf{x}, \mathbf{y}\}$, and $\mathbf{x}$ denotes a training input; $\mathbf{y}$ is the desired output of the corresponding input. The goal of the supervised model is to learn from a sequence of data pairs, in order to minimize the difference between the model outputs and the desired outputs following some standard updating algorithms, e.g. least mean square (LMS). After the function is optimized, the model is able to predict a output $\mathbf{y}$ for a given new input $\mathbf{x}$. If the output is a continuous value, then we are dealing with a regression problem; if the output gives a class label, then it is a classification problem. Moreover the function should be generalizable so that it is also able to predict unseen situations. The mechanism of supervised learning is consistent with *concept learning* in human psychology. Concepts are the mental categories, which guide us to identify objects based on a set of common relevant features of the concepts. Thus *concept learning* refers to a learning task, where humans train themselves to classify objects by observing a set of example objects along with their class labels, and to simplify what has been observed. The simplified information will then be applied to new objects.

Reinforcement learning is like supervised learning in that the model has a reference to look at, so as to maximize or minimize a certain entity. As mentioned, for supervised learning the reference is the desired outputs, and the model is built to match system outputs with the desired outputs in a certain degree. For reinforcement learning, the model interacts with an environment, and infers actions based on the interaction information. The actions in turn can affect the

environment, and excite it to give out more rewards or less punishments. In this sense, the reference is the rewards or punishments, and the action taken by the model will affect the environment to react in a desired way. However reinforcement learning also differs from supervised learning since the ground truth input and output pairs are unknown, and neither the optimal actions. In short, reinforcement learning intends to find a principle that maps the states of environment to actions, which the model needs to take in those states.

Game theory is under the rubric of applied mathematics. It can be seen as an extension or generalization of reinforcement learning. The interaction between two agents still exists. The difference between two learning schemes lies in the characteristics of environments. Reinforcement learning interacts with a static environment. On the contrary, in game theory the environment is dynamic, and it can consist of models or machines, which also take actions and receive rewards (penalties). Therefore a model's success of taking actions depends on the actions of other models in an environment.

Instead of learning a function, which predicts an output for a given input datum, unsupervised learning typically regards input signals as a set of random variables, and it investigates and extracts patterns in input vectors. Therefore the model has no reference like the target $\mathbf{y}$ in supervised learning, neither it has feedback, rewards or punishments from the environment. The pattern may be shown by different representations of the data. Two classic roles of unsupervised learning are clustering and dimensionality reduction. The clustering property here reflects the unsupervised grouping of data in COCA analysis.

Semi-supervised learning, stated by its name, are in-between supervised and unsupervised learning. It normally uses a small amount of labeled data with a number of unlabeled data. Labeled data are usually expensive, and require experts in the area of the particular learning problem (e.g. music genre labels) to manually tag samples with labels. Instead, unlabeled data are comparably easy to access and less expensive. Hence semi-supervised learning is more practical. For detailed survey on semi-supervised learning, see [88].

So far we have touched upon several sub-fields of machine learning, and the brief introduction was based on the comparison among them. Hereafter we focus on unsupervised learning starting with the probabilistic modeling and information theory.

4.2 Unsupervised Learning

Unsupervised learning extracts patterns from random variables, and the patterns can be statistical regularities of inputs [7, 69]. Basically learning statistical properties can be fulfilled by means of learning probabilistic models of input data $P(\mathbf{X})$. Here $\mathbf{X}$ consists of a sequence of data points: $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, and each datum may have multi-dimensions $\mathbf{x}_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,d}]^T$. To keep the consistency, we will use $\mathbf{X}$ to represent a data matrix, and $\mathbf{x}$ for a single datum. Data are regarded as independent and identical distributed (iid) samples extracted from a particular distribution. Therefore

$$P(\mathbf{X}) = \prod_{i=1}^{n} P(\mathbf{x}_i). \tag{4.1}$$

4.2.1 Probability in Information Theory

Probability has two different definitions. Before the prevalence of Bayes' theorem, probability only represents the frequency of random variables in random experiments, and a classical example is 'coin toss'. Another view of probability suggests the degree of belief where no random variables are involved. Probability and the degree of belief can be equalized if they both satisfy *Cox axioms* [52].

The Bayes rule suggests that inferences are based on assumptions, and probabilities are used to describe assumptions:

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{\sum p(\mathbf{x}|\theta)p(\theta)}, \tag{4.2}$$

where $p(\theta)$ is the prior probability of θ; $p(\mathbf{x}|\theta)$ is the likelihood of θ, and it is a function of both θ and $\mathbf{x}$. For fixed θ, the likelihood $p(\mathbf{x}|\theta)$ becomes a probability over $\mathbf{x}$. These assumptions make probabilities subjective. However this inference rule will also end up with the same results, if the same assumptions and data are used.

Another way to represent the probability of a random event can be the amount of *self-information*. In information theory, self-information is a measure of the information content associated with a probabilistic event, and is also called *Shannon information content*:

$$\mathrm{h}(\mathbf{x}) = -\log_2 P(\mathbf{x}). \tag{4.3}$$

As shown in Equation 4.3, the larger the probability, the smaller the *self-information*, indicating the information that the random event indeed occurred. The expected value of the *self-information* of an event is defined as the event's entropy:

$$\mathrm{H}(P) = -\sum_{\mathbf{x}} P(\mathbf{x}) \log_2 P(\mathbf{x}). \tag{4.4}$$

Now we have the entropy of a certain distribution $P(\mathbf{x})$. How do we measure the difference between the $P(x)$ and the 'true' distribution of $\mathbf{x}$, let's say $Q(\mathbf{x})$? Relative entropy provides us with a tool to measure the distance of two distributions, which is also called Kullback-Leible (KL) divergence:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{\mathbf{x}} P(\mathbf{x}) \log \frac{P(\mathbf{x})}{Q(\mathbf{x})}. \tag{4.5}$$

It is not so hard to see that $D_{\mathrm{KL}}(P \parallel Q) \neq D_{\mathrm{KL}}(Q \parallel P)$ and $D_{\mathrm{KL}}(P \parallel Q) \geq 0$ based on *Gibb's inequality* [52].

4.2.2 Hidden Variable Models

Many unsupervised learning models share commonalities and a general formulation can be casted, which is named the hidden variable model. A number of unsupervised learning models use this framework with different constraints on variables. The description of hidden variable models is rather introductory, for detailed mathematical description, see [59].

The hidden variable model has the following form:

$$\mathbf{y} = \Lambda \mathbf{x} + \epsilon, \tag{4.6}$$

where in general $\mathbf{y}$ denotes a d-dimensional observation; $\mathbf{x}$ stands for k-dimensional vector corresponding to hidden variables; thus Λ has dimension d-by-k ($k < d$), and is a mixing matrix which maps hidden space into data space; and ϵ denotes a d-dimensional noise vector. By giving these variables various constraints, a number of linear component analysis methods will emerge, such as principal component analysis (PCA), factor analysis (FA), non-negative matrix factorization (NMF), and independent component analysis (ICA).

Principal Component Analysis has been introduced as a dimensionality reduction tool in Section 3.2.4. Here we are going to describe PCA from a hidden

variable model viewpoint. Probabilistic PCA [78] is derived from the general hidden variable model, shown in Equation 4.6, with the following restrictions:

1. Firstly, $\mathbf{x}$ is assumed multivariate Gaussian distributed with $\mathcal{N}(0, I)$, I represents an identity matrix;
2. Secondly, ϵ is d-dimensional zero-mean isotropic multivariate Gaussian noise, i.e. $\mathcal{N}(0, \Psi = \epsilon\epsilon^{\intercal} = \sigma^2 I)$.

Since Gaussian distribution is invariant to linear transformations, Gaussian distributed data will keep the problem staying in the Gaussian domain. Thus the observation $\mathbf{y}$ is still Gaussian distributed with $\mathcal{N}(0, \Sigma)$ where $\Sigma = \Lambda\Lambda^{\intercal} + \sigma^2 I$.

If we further limit $\sigma \to 0$ in probabilistic PCA, the conventional PCA is presented, where the covariance of observations is simplified as $\Sigma = \Lambda\Lambda^{\intercal}$. Conventional PCA is a common tool to reduce original data into low dimensional space by projecting data along the principal components. Principal components are the ones possessing most variance of data. Here $\Sigma = \Lambda\Lambda^{\intercal}$ will be maximized with the constraint that Λ is orthogonal, which forces hidden variables to be uncorrelated. Conventional PCA is straightforward, and the eigen-decomposition of the covariance will directly show the solution. Considering the limitation $\sigma \to 0$ in $p(\mathbf{x}|\mathbf{y}, \Lambda, \sigma)$, we will end up with $\mathbf{x} = \Lambda^T \mathbf{y}$.

Factor Analysis is also one of the basic dimensionality reduction forms. It models the covariance structure of multi-dimensional data by expressing correlations in a lower dimensional latent subspace. It formulates the general hidden variable model in a similar way as PCA, except for the noise assumption. The constraint on noise is more relax in FA, and $\varepsilon \sim \mathcal{N}(0, \Psi)$, where Ψ is a diagonal matrix with different entries along the diagonal, meaning noise levels are different among dimensions. Since $k < d$, the multi-dimensional datum $\mathbf{y}$ is transformed into lower dimensions. Here $\mathbf{x}$ is called hidden factors, and correspondingly Λ is called factor loading matrix. Factor analysis aims at estimating Λ and Ψ, in order to give a good approximation of covariance structure of $\mathbf{y}$.

Non-negative Matrix Factorization [46] can also be seen as a latent variable model. It has been introduced in [45] as a method for parts-based object recognition. The decomposition follows Equation 4.6 as well with $\epsilon = 0$. A more general formulation of NMF is in matrix form:

$$\mathbf{Y} \approx \Lambda\mathbf{X}, \tag{4.7}$$

where $\mathbf{Y}$ is d-by-n non-negative matrix, with d dimensions and n samples; this matrix is then approximately factorized into a d-by-r non-negative matrix Λ,

and a r-by-n non-negative matrix $\mathbf{X}$. Λ is regarded as the basis, and $\mathbf{X}$ is the encoding matrix. The r is chosen to satisfy $(d+n) \times r < d \times n$, so that matrix Λ and $\mathbf{X}$ are smaller than original data matrix $\mathbf{Y}$, thence data are compressed.

Lee argued that the factorization of an observation matrix in terms of a relatively small set of cognitive components, each consisting of a non-negative feature vector and a non-negative activation vector, leads to a parts based object representation. The non-uniqueness of components is a major challenge for NMF, and has been discussed in detail in [18]. A possible route to more unique solutions, hence, potentially more interpretable and relevant components is to add a priori knowledge, e.g., in form of independence assumptions.

It is often the case that the hidden variables we are looking for are not always Gaussian distributed. For instance, as illustrated in Chapter 3 data from sensory analysis are usually sparse. ***Independent Component Analysis*** successfully extend the hidden variable model to fit non-Gaussian factors. In ICA, hidden variables are defined as independent non-Gaussian distributed sources.

Independence is a stronger assumption than uncorrelation. It can be illustrated by probability distributions. Let us denote $p(\mathbf{x}, \mathbf{y})$ the joint distribution of two variables $\mathbf{x}$ and $\mathbf{y}$. Hence the marginal distribution of $\mathbf{x}$ and $\mathbf{y}$ are given as:

$$\begin{aligned} \widehat{p}(\mathbf{x}) &= \int p(\mathbf{x}, \mathbf{y}) d\mathbf{y}, \\ \widehat{p}(\mathbf{y}) &= \int p(\mathbf{x}, \mathbf{y}) d\mathbf{y}. \end{aligned} \tag{4.8}$$

Two variables $\mathbf{x}$ and $\mathbf{y}$ are considered independent, if and only if the joint distribution follows the factorization below:

$$p(\mathbf{x}, \mathbf{y}) = \widehat{p}(\mathbf{x})\widehat{p}(\mathbf{y}). \tag{4.9}$$

This factorization of the joint distribution also can be transformed into the following form, which will be compared later with the expression of uncorrelation:

$$\mathrm{E}\{f_1(\mathbf{x})f_2(\mathbf{y})\} = \mathrm{E}\{f_1(\mathbf{x})\}\mathrm{E}\{f_2(\mathbf{y})\}, \tag{4.10}$$

where $\mathrm{E}\{\cdot\}$ denotes Expectation; and $f_1(\cdot)$ and $f_2(\cdot)$ are two functions.

However uncorrelation between two variables $\mathbf{x}$ and $\mathbf{y}$ is represented by their covariance:

$$\begin{aligned} Cov(\mathbf{x}, \mathbf{y}) &= \mathrm{E}\{[\mathbf{x} - \mathrm{E}\{\mathbf{x}\}][\mathbf{y} - \mathrm{E}\{\mathbf{y}\}]\} \\ &= \mathrm{E}\{\mathbf{x}, \mathbf{y}\} - \mathrm{E}\{\mathbf{x}\}\mathrm{E}\{\mathbf{y}\} = 0. \end{aligned} \tag{4.11}$$

Therefore we can say that independence of variables indicates uncorrelation when we define $f_1(\mathbf{x}) = \mathbf{x}$ and $f_2(\mathbf{y}) = \mathbf{y}$, but uncorrelation does not imply independence.

ICA is able to estimate both the mixing matrix Λ and the sources $\mathbf{x}$. This is done by either maximizing the non-Gaussianity of the calculated sources or minimizing the mutual information [36].

PCA, ICA and FA have all been utilized by COCA analysis for different applications. More details of these unsupervised learning models will come along with a variety of topics later. The remainder of this chapter will elaborate on two speech related topics, which use unsupervised learning techniques to look for signatures of phonemes and speaker identities.

4.3 'Fingerprint' of Phonemes

Phonemes are defined as the class of sounds that are consistently perceived as representing a certain minimal linguistic unit in [16]. However phonologists have different views of phonemes, and two major ones are: in the American structuralist tradition, a phoneme is defined according to its allophones and environments; in the generative tradition, a phoneme is defined as a set of distinctive features [50]. An allophone is a phonetic variant of a phoneme in a particular language. According to the first view, the same phoneme can sound slightly different in different languages and environments. In American English approximately 40 phonemes are in use, of which 12 are vowels. Vowels vary in temporal duration between $40 - 400$ $msec$ [16].

Phonemes in average can be described by short-time features. It only involves the presentation of phonetic units, neither the information of a whole word, nor the semantics. Thus phoneme relevant cognitive function is on quite a low level. In appendix B phonemes have been studied in a relatively small data set. The unsupervised learning method, PCA has translated the original data into a lower dimensional subspace based on the orthogonal basis vectors derived from SVD. Four letters 's', 'o', 'f' and 'a' were collected from TIMIT letter dataset, which includes two trials of clean speech of 26 letters pronounced by one person. Following the preprocessing procedure, features were found distributed in a sparse linear mixture manner, see Figure 2, 3 in appendix B. Since phoneme information can be carried by short-time frames, feature integration were excluded. Cognitive components of phoneme /e/ opening 's' and 'f' were identified. We speculate that these phoneme-relevant cognitive components contribute towards the well-known basic 'invariant cue' characteristics of speech [10].

The theory of acoustic invariants indicates that perceived signals are derived as stable phonetic features, despite of different acoustic properties produced by different trials from one speaker, and by different speakers. Moreover Damper

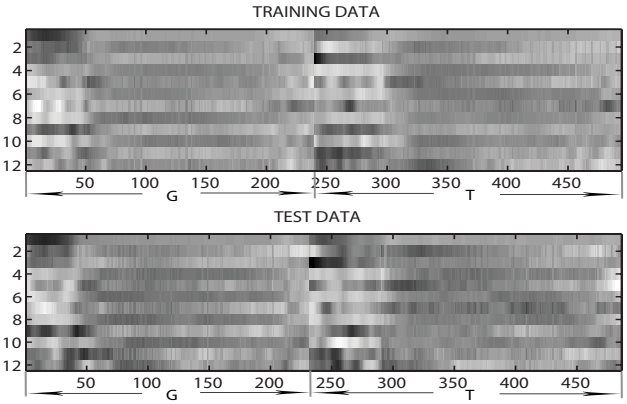

Figure 4.1: Original MFCCs. It shows the temporal development of 12-dimensional MFCCs from both training and test set of letter 'g' and 't', which both ends with vowel: /i:/.

has shown that although speech signals may vary due to coarticulation, the relation between key features follows a consistent and invariant form [14].

4.3.1 'Invariant Cue' of Single Speaker

To further confirm our previous finding on 'invariant cue', we will carry out a number of experiments to test the 'invariant cue' theory. We start with a simple low-level COCA experiment on one person's speech. By studying the structure of features extracted from one particular phoneme in the space, which consists of different phonemes pronounced by one person, we look at the variance of this particular phoneme. In other words, we study the phoneme included in different words or letters, and investigate the grouping outcome by unsupervised learning method.

The chosen phoneme is included in two letters: 'g' and 't', and their phonetic symbols are: /dgi:/ and /ti:/. Theoretically they share a same phoneme: vowel /i:/. According to the 'invariant cue' theory, phonetic features derived from different words are invariant to their environments (here we mean the surrounding phonetic features, e.g. /dg/ and /t/) and different trials.

We used letter 'g' and 't' from TIMIT letter database. The first trials of these two letters were used as training set, while the second as test set. 12-dimensional

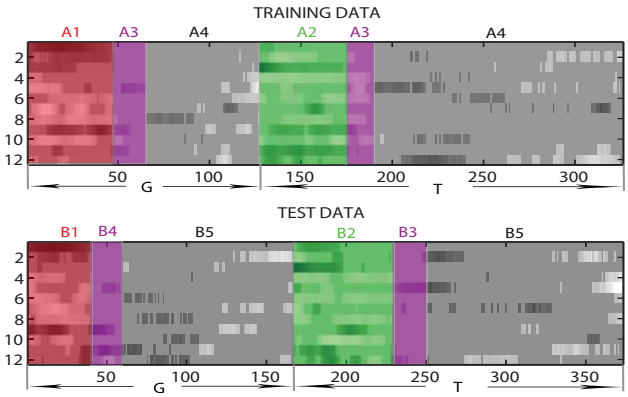

Figure 4.2: Sparsified MFCCs. It shows the sparsified MFCC coefficients with threshold $z = 1.2$, which keeps 68% energy. Both training and test sets are given. Several regions are market corresponding to different phonetic groups. Their locations in the scatter plots are shown in Figure 4.3 and 4.4.

MFCCs were extracted from 333 samples at 10 kHz with 95% overlap between adjacent short-time frames. These MFCCs used hamming windows in time domain and triangular mel-filters in the absolute log frequency domain. All the MFCCs will be extracted under this setup with different frame length in the upcoming experiments, unless otherwise specified. MFCCs from two letters were concatenated. Threshold for sparsification was set to keep 68% of total energy in the remaining coefficients. Figure 4.1 shows the temporal development of 12-dimensional MFCCs for both training and test sets. Boundaries of two letters are obvious, which is 240 for training set and 230 for test set. The transient of phonemes within each letter is also remarkable. We show the sparsified MFCCs in Figure 4.2, note that samples obtaining zero energy have been removed. Afterwards, SVD found eigenvectors, and hence features were projected along principal components. Figure 4.3 gives the scatter plot of the retaining training set features in the subspace of two principal components. The 'ray-structure' is striking. We have studied the samples in the small data set, and found out their corresponding locating areas in the temporal development of MFCCs. The second plot in Figure 4.3 are divided into 4 regions, marked from A1 to A4. Since the scatter plot is in 2 dimensions, regions indicating different phonetic features could have overlap. Their area coverage has been roughly marked in the time domain, shown in the upper panel of Figure 4.2. Loosely speaking, region A1 indicates phonetic features related to /dg/; A2 corresponds to /t/; A3 are the transient parts from both /dg/ to /i:/ and /t/ to /i:/; while region A4 is the 'invariant cue' we are looking for: the vowel /i:/ existing in both

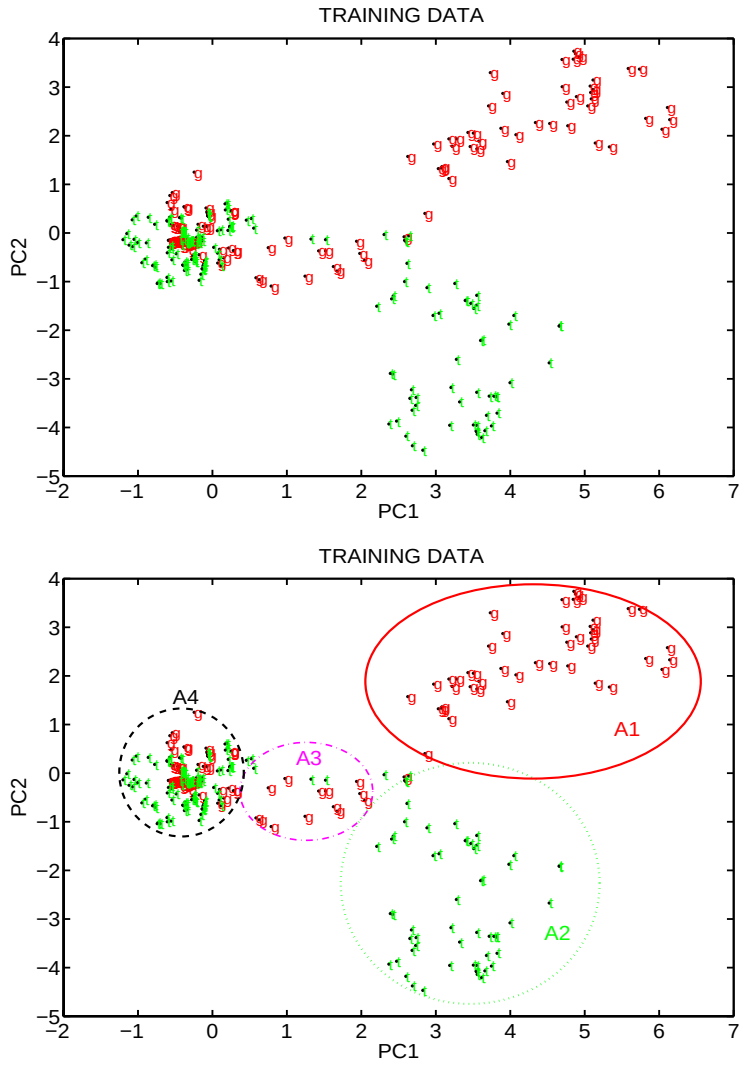

Figure 4.3: Scatter plot of training set MFCCs extracted from letter 'g' and 't' in the subspace of first two principal components. Samples labeled with g and t indicate their affiliations. The 'ray-structure' is observable. By studying temporal locations of these samples, we allocate them as regions in the sparsified MFCCs (Figure 4.2). Regions A1 represents the starting part of letter 'g': phoneme /dg/; A2 represents the beginning of 't': phoneme /t/; and A3 is the transients between phonemes within one letter; and A4 denotes the area of /i:/ sound ending both letters, which indicates the so-called 'invariant cue'.

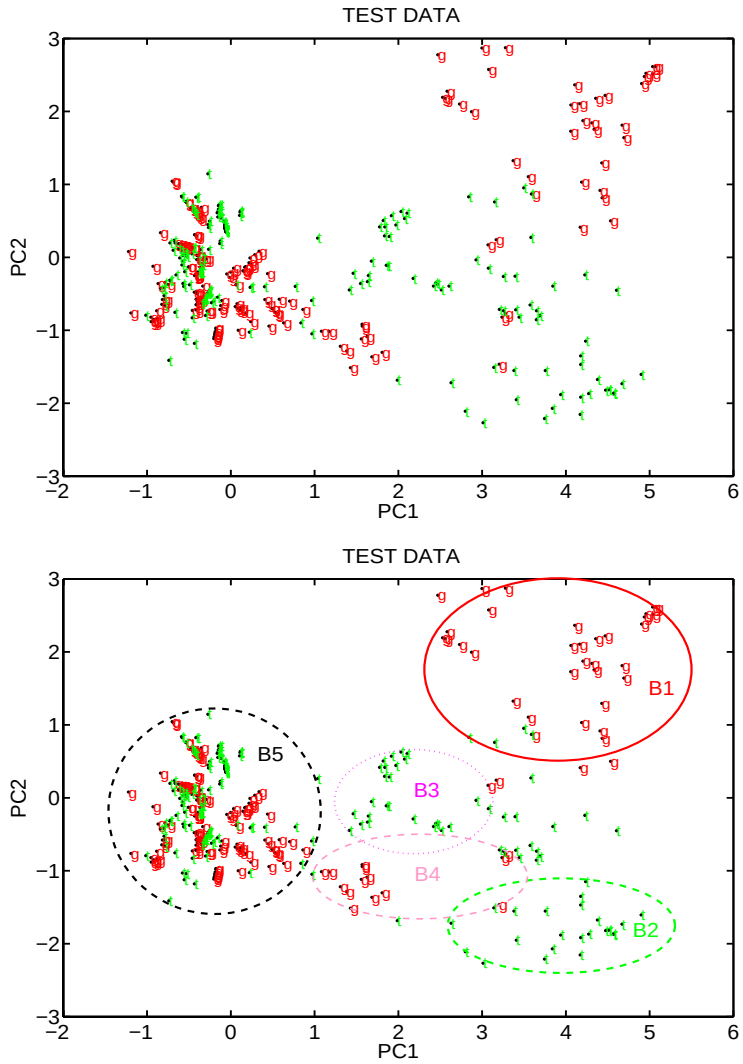

Figure 4.4: Scatter plot of test set MFCCs extracted from letter 'g' and 't' in the subspace. Samples labeled with g and t. The 'ray-structure' is revealed. The temporal locations of samples are categorized into 5 regions. Regions B1 represents the starting part of letter 'g': /dg/; B2 represents the beginning of 't': /t/; and B3 and B4 are the transients between phonemes within one letter; and B5 indicates the region of the /i:/ sound from 'g' and 't'.

letters 'g' and 't'. Results coincide with those included in appendix B, and also the LSI analysis on text data, which shows that text having similar semantic meanings locates near one another in the semantic space. This conclusion can be translated, in the sense of phonemes, to such that samples sharing similar phonetic characteristics locate close to each other in the phonetic space.

To test the generality of findings, we preprocessed test set in the same procedure as training set, and projected sparsified MFCCs along the eigenvectors of the training set. 'Ray-structure' has also been found, shown in Figure 4.4. Same as Figure 4.3, the upper panel is the scatter plot of test set in the first two principal components; and the second panel is the same plot marked with regions: from B1 to B5, indicating different phonetic units. Their corresponding locations are roughly drawn in the lower panel of Figure 4.2. Similar to training set regions, B1 is the group of MFCC samples from phoneme /dg/; B2 refers to /t/; B3 is the transient components from /t/ to /i:/; while B4 is the transient from /dg/ to /i:/; finally B5 represents the common /i:/ sound from both 'g' and 't'.

4.3.2 'Invariant Cue' of Multi-Speaker

So far the so-called 'invariant cue' has been discovered in a simple situation, where only one speaker's speech has been studied. Based on the general definition, stable phonetic features should also exist across speakers, despite of variations from individuals. This subsection will focus on revealing 'invariant cue' from multi speakers.

We select data from TIMIT database [27]. TIMIT collects reading speech from 630 native American English speakers. Each speaker reads 10 sentences in total, and each sentence lasts approximately $3s$. The phoneme transcription is available. Here we will study the letter 't' sound from three speakers: two male speakers and one female speaker. We truncate the /ti:/ part from the word 'teeth'. This word is covered by sentence SX333 from speaker FCJF0, and sentence SI648 from speaker MDPK0. The last speaker is from TIMIT letter database. Therefore the basic speech information from each speaker is /t/ + /i:/, but they are extracted from different environments. The feature extraction setup was similar to the previous experiments. Figure 4.5 provides the original 12-dimensional MFCCs and the sparsified ones, and zero energy samples are removed. Boundaries among three speakers: M1, F1(FCJF0) and M2 (MDPK0) are quite obvious. The lower panel is tagged by regions (C1 to C4) corresponding to the regions defined in the scatter plots: Figure 4.6. In the case, the sparse linear mixture is revealed again. Region C1, C2 and C3 show the phoneme /i:/ from all three speakers; while as region C4 includes phoneme /t/. Here we notice that the 'ray-structure' is not a one-to-one case: one ray

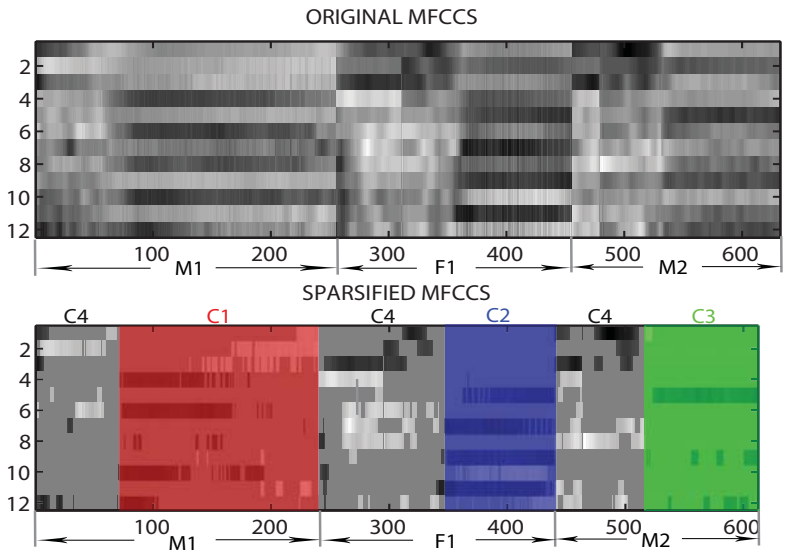

Figure 4.5: Original and sparsified MFCCs of sound /ti:/ from 3-speaker training set. It shows the temporal development of 12-dimensional MFCCs, which share the same phonemes /t/ and /i:/. M1, F1 and M2 denote speaker ID. Regions from C1 to C4 refer to the groups of some particular phonetic features, and locations of these samples in the subspace are shown in Figure 4.6. Since /ti:/ sound was extracted from word 'teeth' for speaker F1 and M2 (only speaker M1 pronounced letter 't'), their temporal development of MFCCs look alike.

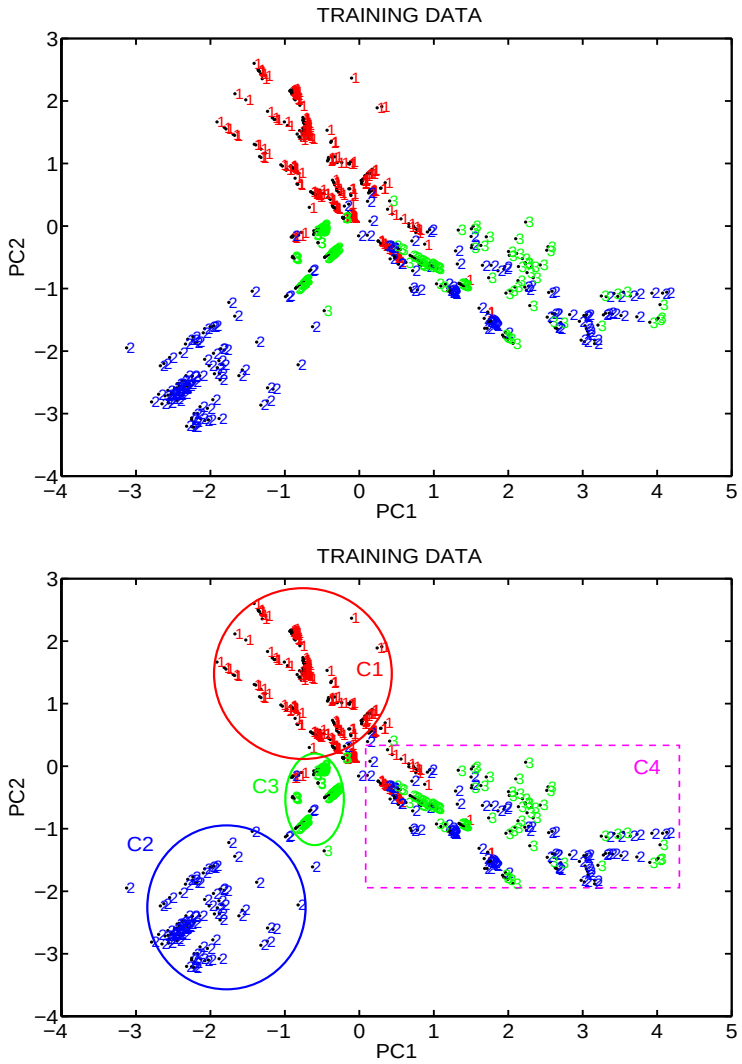

Figure 4.6: Scatter plots of MFCCs extracted from /ti:/ sound in the subspace formed by first two principal components. The 'ray-structure' is striking. Samples labeled with 1,2 and 3 are from three speakers: M1, F1, M2 respectively. The lower panel is labeled with regions C1 to C4, where C1 to C3 represent /i:/ from three speakers, and C4 is the /t/ sound pronounced by all of them. Region C2 and C3 seem to follow the same 'rays' emanating from the origin with different amplitudes.

to one phoneme (see phoneme /i:/), neither it is a one ray to one speaker case (see speaker F1 & M2). However similar phonetic features do tend to group, if we divide the 2D space into left and right from around $x = 0.2$, features of phoneme /i:/ from all three speakers locate in the left side, while as phoneme /t/ in the right side. Moreover, if we have a close look at the region C2 and C3, it is obvious that the /i:/ sound from speaker F1 & M2 locate along the same 'rays' with different amplitudes. It can be also observed from Figure 4.5 that MFCCs from these two speakers do share large similarities comparing to speaker M1, due to the fact that they were both extracted from word 'teeth', and the letter 't' was pronounced only by M1. Besides the 'ray-structure' of phoneme /t/ shared by three speakers, a more complicated cue /i:/ has been revealed, all of which support the theory of 'invariant cue' from different angles. Further more Figure 4.6 indicates some speaker-specific properties as well. By looking at sample locations of male and female speakers, F1 has part of the data locating apart from features of M1 and M2, which may imply that our chosen features and COCA model, are capable of specifying individuals. More details will come in Section 4.4.

Invariant cues are hard to identify, and researchers on this subject have different views: some believe that phoneme is the fundamental unit in speech perception, and invariant characteristics are derived from these units; some believe that invariant cues are not static but dynamic, and therefore can not be associated with a single phoneme; some even question about the phoneme being the fundamental unit, and claim that the unit in speech perception is not unique, but is dependent on the focus of attention of the brain. One of the reasons is that they believe speech perception is based on syllables or words, and hence phonemes are not perceived, but perhaps inferred from the perceived syllables or words [19]. Nevertheless, the 'invariant cue' has been revealed here based on the first view that invariant property is derived from phonemes.

4.3.3 Independent Components of Phonemes

LSA-like PCA found us 'ray structure' of phoneme features. However whether the generalizable structure can assist phoneme recognition in general, still needs to be explored. Appendix C has demonstrated that applying ICA to group phoneme features is more appropriate than only applying PCA, which is too constrained or in some instances too flexible. To keep it consistent, we here show a group of results provided by ICA models on data introduced in subsection 4.3.1: one speaker's voice pronouncing letter 'g' and 't'.

Maximum likelihood ICA algorithm with 6 independent components has been used on PCA coefficients representing 'g' and 't' in the mel-cepstrum domain.

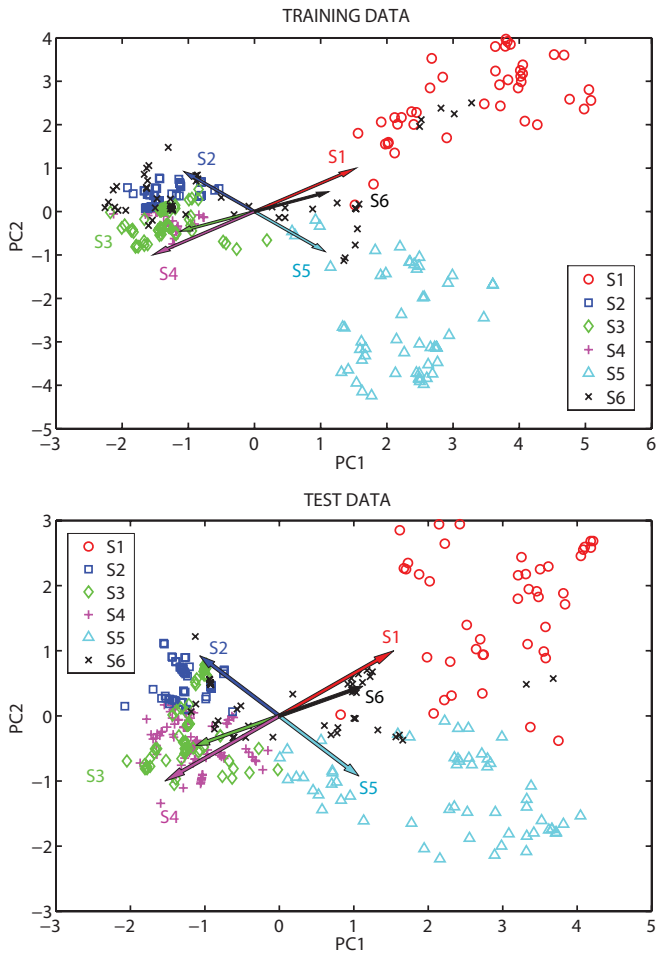

Figure 4.7: Scatter plots of 6 ICA components in the subspace of the first two principal components. Data are tagged based on the recovered independent sources, and arrows are the column vectors of the mixing matrix.

Following Equation 4.6 with $\epsilon = 0$, Λ has been recovered and consequently sources $\mathbf{x} = \Lambda^{-1}\mathbf{y}$. The recovered sources of both training and test sets are shown in Figure 4.7. Columns of Λ are regarded as vectors to direct the separation of independent sources. Similar to the representation shown in Figure 4 and 5 in appendix C, temporal locations of samples belonging to each source are given as vertical lines in different colors in Figure 4.8, for both training and test data separately. It proves that ICA is capable of recovering independent sources: the first source (red lines) represents phoneme /dg/; the majority of the fifth source (cyan lines) represents phonetic unit /t/; and the rest sources (2^{nd}, 3^{rd}, 4^{th} and 6^{th}) together account for the shared /i:/ sound by letter 'g' and 't'. In the meanwhile we noticed that part of the third source (green lines) occurs at the transient areas: the transforming parts between phonemes. Whether the transient part can be represented by single- or multi-independent components still needs to be further explored. The classification rate by hard assigning each sample to one source is about 95.1% for training data and 89.5% for test data.

4.4 'Voiceprint' of Speakers

Speaker recognition is one of the speech based engineering applications, and it involves two applications: speaker identification and speaker verification. This technique makes it possible to use speakers' voice to detect their identities, and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers.

4.4.1 Introduction of Speaker Recognition

Speaker verification is the process of determining whether the speaker identity is who the person claims to be. It performs a one-to-one comparison (binary decision) between the features of an input voice and those of the claimed voice, which is registered in the system. This comparison is often called pattern matching, and if the match is above a certain threshold, the claimed identity is verified. Using a high threshold, system gets high safety and prevents impostors to be accepted, but at the same time takes the risk of rejecting the genuine person, vice versa.

Speaker identification is the process of identifying the ID of a speaker by comparing his/her voice with voices of registered speakers in the database. It is a one-to-many (M) comparison. M speaker models are scored in parallel, and the

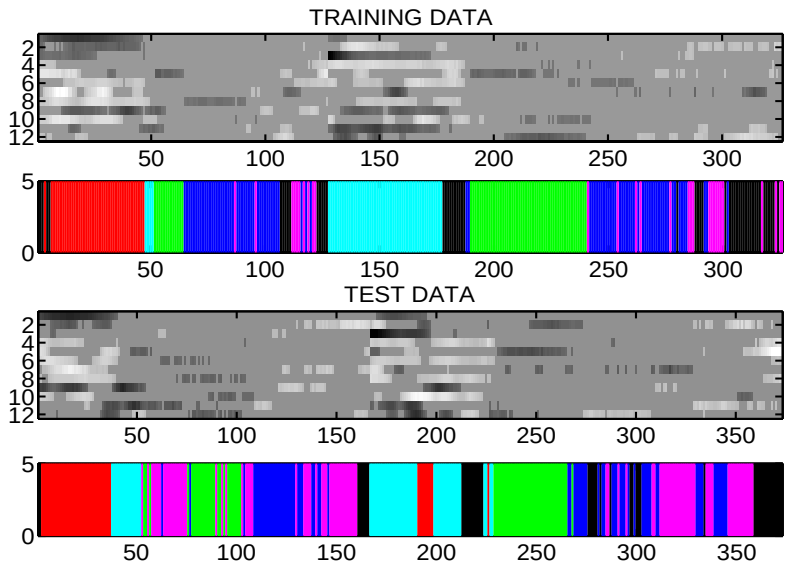

Figure 4.8: Sparsified MFCCs of letter 'g' and 't', and the locations of 6 independent sources. Samples are hard assigned to one source having the largest mixing proportion. Vertical lines with the same color indicate the samples belonging to the same source. Source 1 to 6 are colored as red, blue, green, magenta, cyan, black in the same sequence.

one with highest score is reported if it is a closed-set case. Closed-set means that the unknown voice must come from a set of known speakers; on the contrary open-set means unknown voice may come from unregistered speakers, in which case a 'none of the above' option can be added to the identification system. Based on speech modalities, speaker recognition can be categorized as text-dependent and text-independent situations.

The first speaker recognition machine using spectrogram of voices was invented in the 1960's. It was called voiceprint analysis or visible speech. The voiceprint is the acoustic spectrum of voice, and it has similar definition as fingerprint. Voiceprint analysis could not perform automatic recognition, and human's manual determination is needed. So far a number of feature extraction techniques developed for speech recognition, have been used to serve speaker recognition systems. Since the mid-1980s, speaker recognition field has been steadily getting mature, and commercial applications have been increasing. Features with various representations have been derived, some are calculated in time domain, some in frequency domain [73], and some in both domains [13]. Furthermore, a variety of models can be found for establishing speaker recognition systems, such as Gaussian Mixture Model (GMM) [72] and Hidden Markov Model (HMM) [54], which are the state-of-the-art models in this field. The system in [72] has been frequently quoted. It uses Mel-scale cepstral coefficients (MFCCs). Based on [72], some modifications have been done. In [62] MFCCs were transformed to compensate noise components in the audio channel, and then formants features were calculated and used in classification. In [77], PCA was utilized on the features mentioned in [72]. MPEG-7 as a new technique is used for speaker recognition. MPEG-7, formally named 'Multimedia Content Description Interface', is a standard for describing the multimedia content data that supports some degree of interpretation of the information's meaning, which can be passed onto, or accessed by, a device or a computer code. In [39] MPEG-7 Audio standard were used for speaker recognition problem. MPEG-7 Audio standard comprises descriptors and description schemes. They are divided into two classes: generic low-level tools and application-specific tools. There are 17 low-level audio descriptors (LLD). [39] used a method of projection onto a low-dimensional subspace via reduced-rank spectral basis functions to extract speech features. Two LLDs were used (Audio Spectrum Basis Type and Audio Spectrum Projection Type) together with ICA model to discover speaker identities: for a small set the accuracy was up to 91.2%; for a large set it was 93.6%; and the gender recognition accuracy for small set was 100%.

4.4.2 Cognitive Components of Speakers from Different Text

Human can effortless recognize speaker identities of acquaintances from a short period of speech (e.g. several seconds). Here we envision that speaker relevant cognitive components can be revealed by COCA analysis. We would like to start with a simpler case: speaker-specific characteristics extracted from different text. Therefore the results will not be interfered by the phoneme-like phenomenon illustrated in the Section 4.3. In the following subsection, a more complicated case: speaker-specific characteristics extracted from the same text will be described.

Several experiments have been done with short-time MFCC features, followed by PCA. However no distinguished structures have been found. We hypothesize that speaker identity may lie on a higher cognitive level than phonemes. Based on the preprocessing pipeline introduced in Section 3.2, we extracted 12-dimensional MFCCs at the basic time scale ($20ms$). Whereafter we stacked several frames to construct long feature vectors at a longer time scale, i.e. 1 *sec*, and the overlap between two long frames was 50%. Since we need longer speech signals from each speaker, our in-house speech database: ELSDSR [23] was used. It records around 100 *sec* speech on average for each enrolled speaker, and totally 22 speakers were enrolled. This database is divided into two sets: the recommended training set and test set. The training set is the same for every speaker, and it covers seven paragraphs with 11 sentences in total. In test set each speaker read two sentences, and 44 different sentences were recorded. The recordings of two female speakers (denoted F1 and F2), and one male speaker (denoted M1) were collected for this experiment. Since we were aiming at different text, we carefully chose different text content for speakers. We collected around 32 *sec* training data for each speaker, and 20 *sec* test data per speaker. EBS has discarded most of the coefficients, and only the upper 4% was kept. Figure 4.9 demonstrates the resulting 'ray structure' of training and test set, derived from PCA projection onto the 2^{nd} and 4^{th} principal components. Sparse components for each individual speaker are evident, and 'rays' locate very much separately in the subspace. It is a bit hard to see the structure for speaker F1, since the data are in a smaller scale. A zoomed-in plot and more results can be found in appendix B.

4.4.3 Cognitive Components of Speakers from Same Text

In this experiment we kept text content the same to every speaker, and see how it influences the representations of speaker identity. Another set of speakers have been chosen, still with two females and one male speaker denoted the

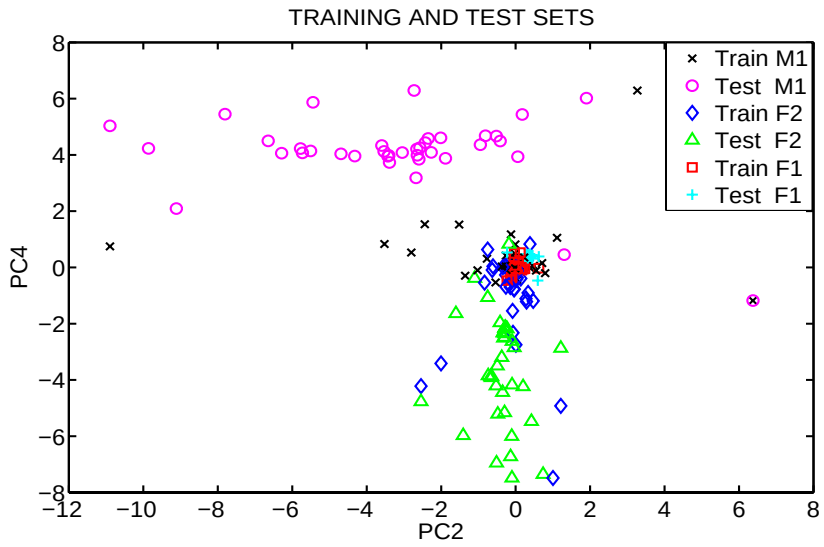

Figure 4.9: Scatter plot of 3 speaker-specific components found from different text. For detailed results, see Fig. 5 in appendix B.

same way. First, MFCCs were extracted from 20 *ms*, and later long vectors representing time scale of 1 *sec* were constructed the same way as before. The training set lasts 52.5 *sec* long per speaker, while as test set lasts 35.5 *sec* long. 7.3% energy was survived from EBS. The scatter plots of training and test sets are shown separately in Figure 4.10. At the first glance, due to the same-text training and test sets show different patterns: training data from three speakers have large overlaps around origin of the coordinate system; while as 'rays' of test data tend to extend along a similar direction. However a close depiction of the data scatter for each speaker individually, elucidates that the training and test data follow a similar scatter tendency with offsets, see Figure 4.11. In short, phoneme-like structures showed up because of the influence brought by the same text content; moreover speaker-dependent structures have also been revealed. We stipulate that this phenomenon is the interaction between the text content and the speaker identity, which echoed the findings in the previously discussed experiment on 'invariant cue' of multi-speaker.

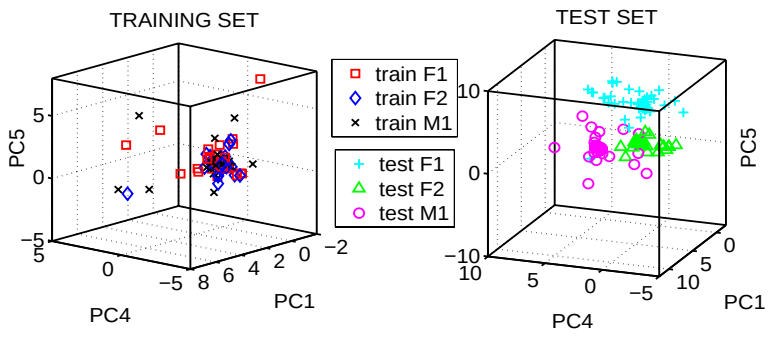

Figure 4.10: Cognitive components of speakers from the same text content. Figure shows the scatter plots of training and test sets separately for 3 speakers. Phoneme-like phenomenon is found, due to the different scatter patterns between training data and test data.

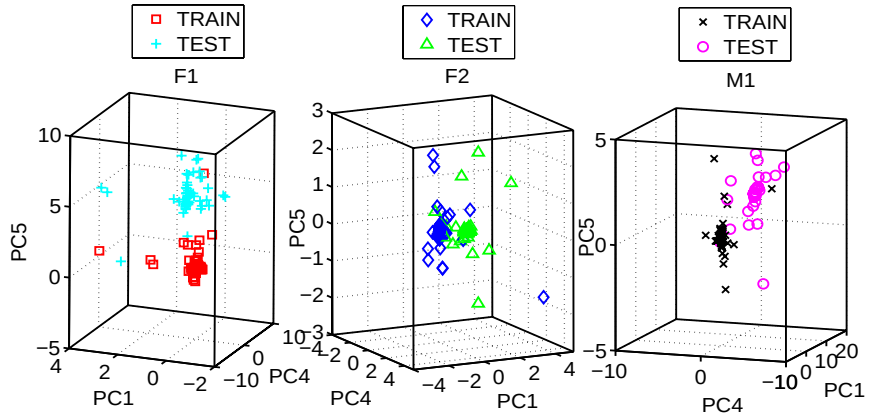

Figure 4.11: Cognitive components of speakers from the same text content. Figure shows the scatter plots of training and test sets for each speaker. The coordinate systems are rotated so as to show the data scatter tendency for individuals. Speaker-specific structures are revealed by investigating the training and test data from each speaker individually: they do follow the same tendency.

4.5 Summary

This chapter started with a brief introduction of machine learning. The introduction was parted into a number of learning techniques based on one taxonomy of machine learning, where unsupervised learning gained more emphasis. Unsupervised learning was further described from the point of view of probabilistic modeling and information theory, and summarized as a hidden variable model, which is a general form of unsupervised linear model shared by PCA, ICA, FA and NMF, etc. with different constraints on variables.

Unsupervised grouping of perceptual data was studied on speech data to discover phoneme and speaker identity relevant statistical regularities. For the study on phoneme data, we first discussed our findings in appendix B, which caught our attention on the issue: 'invariant cue'. To prove the generality of the phenomenon that similar phonetic features group together in the subspace defined by PCA, e.g. /e/ opens both letter 's' and 'f', we constructed two experiments in different conditions: 1) we tried to discover the structure of the same phoneme embodied by different letters pronounced by one speaker; 2) in a more complicated case, we aimed at looking for the same phenomenon across multi speakers. In both setup, we did found 'invariant cue' by PCA. These results are covered by appendix H. ICA proved its capability of discovering independent sources from the first experiment, and with 6 independent components the classification accuracy was around 90% by hard assigning each datum to the most likely source.

Whereafter speaker recognition was introduced, which led to the findings included in appendix B. Speaker-specific cognitive components were found at a longer time scale from different text among 3 speakers; and also from the same text, where a complex finding revealed the interaction between the phoneme-like effect and speakers' 'voiceprint' phenomenon.

The fact that we have found cognitive components of speech at different time scales, motivates our following work in Chapter 5.

Chapter 5

On High-level Cognitive Component Analysis

The definition of COCA has been introduced in Chapter 3. We revisit it here for convenience: COCA is defined as the process of unsupervised grouping of generic data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity. Unsupervised grouping of data to discover their statistical properties has been demonstrated on speech signals in Chapter 4. Here comes the time to test our hypothesis and measure the correlations between the unsupervised grouping results and human cognitive performance.

In this chapter, we represent human cognition by supervised classification of manually labeled data, since labels reflect human cognitive activity in perception, decision making and judgement. Thence the comparison between statistical regularities (discovered by unsupervised grouping of data) and human cognition, turns out to be between the unsupervised learning of data and supervised learning of the same data with labels. Further we will look into some higher level cognitive functions involving e.g. age estimation. We will use the designed protocol to test whether the optimal ICA method is invoked by human cognition on higher level functions. Two sets of unsupervised and supervised models have been suggested. The first set is based on mixture of factor analyzers model. Models have been modified to accommodate our data representations, and the resulting models can be interpreted as ICA-like density models. They

have appeared in appendix E. The design of the second set is more explicit, so as to be directly in line with the COCA independent hypothesis. This set of models apply Bayes' theorem, and have been used in appendix F, G and H.

5.1 ICA-like Density Model

For comparison purpose, a pair of models needs to be chosen to represent the unsupervised grouping scheme and human cognition separately. To keep the comparability, models are preferred to share similar structure. In addition, due to sparse independent representations of data, models should be able to accommodate the sparse linear 'ray structure'. The unsupervised and supervised models are designed on the same base to form an ICA-like density model. This can be achieved by modifying an existing popular model, namely mixture of factor analyzers (MFA). The modification follows ideas represented in *Soft-LOST* and *Hard-LOST* (Line Orientation Separation Technique) models [65, 64]. To have a clear understanding of model structures, let us open up this section with the *LOST* models.

5.1.1 *LOST*

In 2004, two line oriented separation models have been introduced, based on a same standing point. One is *Hard-LOST*, a method derived from k-means algorithm to cope with sparse linear mixing data. Another is *Soft-LOST*, invoking Expectation-Maximization (EM) algorithm to define a mixture of oriented lines. Both models are intended to solve blind source separation problem in even-determined and under-determined cases. Here we will follow the line of *Soft-LOST* to introduce these two models. Our resulting ICA-like density model adopts the ideas from both *LOST* models, and applies them to MFA model.

In the context of linear source separation, the model can be represented as a hidden variable model introduced in Section 4.2. We rewrite Equation 4.6 here:

$$\mathbf{y(t)} = \Lambda\mathbf{x(t)} + \epsilon(t). \tag{5.1}$$

The dimension of observation $\mathbf{y}(t)$ at time t is determined by the number of sensors d. The dimension k of $\mathbf{x}(t)$ implies the number of hidden sources. Thus

$\mathbf{y}(t) = [y_1(t), y_2(t), \ldots, y_d(t)]^T$, and $\mathbf{x}(t) = [x_1(t), x_2(t), \ldots, x_k(t)]^T$. The dimensions of $\mathbf{x}(t)$ are mixed in a linear way to form $\mathbf{y}(t)$. In the case that $\epsilon \approx 0$, it becomes a typical noise free ICA model when $\mathbf{x(t)}$ is assumed non-Gaussian. We can regard each dimension of $\mathbf{X}$ as a set of one-dimensional data, and $\mathbf{Y}$ as mixture of these data projected into a d-dimensional space. The columns of mixing matrix Λ give orientations for projection.

Instead of regarding Λ matrix as a whole, O'grady and Pearlmutter see its columns separately as oriented lines $\mathbf{v}_i$. Accordingly the problem has been converted to looking for line orientations, and Λ can be estimated as:

$$\widehat{\Lambda} = [\mathbf{v}_1] \ldots [\mathbf{v}_k]. \tag{5.2}$$

The prefix '*Soft*' or '*Hard*' indicates the assignment type used to assign data to lines. In the beginning, the model randomly selects k line orientation vectors $\mathbf{v}_i$. *Soft-LOST* afterwards applies EM algorithm to update these $\mathbf{v}_i$. In **E-step**, all data points $\mathbf{d}_1, \ldots, \mathbf{d}_T$ are partially assigned to each line orientation following:

$$\begin{aligned} z_{ij} &= \|\mathbf{d}_j - (\mathbf{v}_i \cdot \mathbf{d}_j)\mathbf{v}_i\|^2 \\ \widehat{z}_{ij} &= \frac{exp(-\beta z_{ij})}{\sum_{i'} exp(-\beta z_{i'j})}, \end{aligned} \tag{5.3}$$

where β is the softness parameter, which controls the boundaries of regions attributed to each line vector. $\widehat{z}_{ij}$ gives how much each data point j contributes to each line i, thus $\sum_i \widehat{z}_{ij} = 1$. In **M-step**, the covariance matrix of weighted data for each line are calculated as:

$$\Sigma_i = \frac{\sum_j \widehat{z}_{ij} \mathbf{d}_j \mathbf{d}_j^T}{\sum_j \widehat{z}_{ij}}. \tag{5.4}$$

The vectors $\mathbf{v}_i$ are then updated to be the first eigenvectors of the covariance matrices, i.e. the eigenvector with the largest eigenvalue. The EM steps will be carried out until convergence, and the final line orientations are adjoined according to Equation 5.2 to form the estimated $\widehat{\Lambda}$.

However *Hard-LOST* utilizes *winner-takes-all* strategy, and in this case $\beta \to \infty$ and $\widehat{z}_{ij}$ is either 1 or 0. 1 indicates that a data point j is assigned to the line i since $z_{ij} < z_{i'j}, (i \neq i')$. The stochastic gradient algorithm (SGA) with constraint $\|\mathbf{v}_i\| = 1$, will help the model to determine line orientations. SGAs

are run independently for each line orientation through out all data points, which are hard-assigned to this line, and the final results are used to form matrix $\widehat{\Lambda}$. In short, *Hard-LOST* is derived from the classic k-means method, where the k cluster centers are replaced by k line orientation vectors, and the distances from data points to line vectors substitute the distances to cluster centers.

5.1.2 Unsupervised ICA-like MFA

Factor analysis is one of the basic dimensionality reduction forms. It models the covariance structure of multi-dimensional data, and expresses correlations in a lower dimensional latent subspace. For mathematical formula of FA and variable constraints, see subsection 4.2.2. While the FA is globally linear and Gaussian, we can model non-linear non-Gaussian processes by invoking a so-called mixture of factor analyzers (MFA):

$$\begin{aligned} p(\mathbf{y}) &= \sum_{\mathbf{x}} p(\mathbf{y}, \mathbf{x}) \\ &= \sum_{\mathbf{x}} p(\mathbf{y}|\mathbf{x}) p(\mathbf{x}) \\ &= \sum_{i=1}^{m} \sum_{\mathbf{x}} p(\mathbf{y}|\mathbf{x}) p(\mathbf{x}|i) p(i), \end{aligned} \tag{5.5}$$

where $\mathbf{y}$ is an observation; $\mathbf{x}$ is hidden factors; $p(i)$ and m are mixing proportions and the number of factor analyzers. Factors in each factor analyzer are still assumed Gaussian distributed: $p(\mathbf{x}|i) = p(\mathbf{x}) \sim \mathcal{N}(0, I)$; and noise is uncorrelated Gaussian distributed $\mathcal{N}(\mathbf{u}|0, \Psi)$ with a diagonal matrix Ψ. MFA combines factor analysis and the mixture of Gaussians model (MoG), and hence can simultaneously perform clustering, and dimensionality reduction within each cluster, see [28] for a detailed review.

To meet our request for unsupervised learning of sparse data representations, MFA is modified to form an ICA-like line based density model, and the modification borrows ideas from both *Soft-LOST* and *Hard-LOST* models. We adopt EM procedure. In **E-step**, we calculate the log posterior probability $\log p(i|\mathbf{y})$ according to the Bayes' theorem:

$$\begin{aligned} \log p(i|\mathbf{y}) &= \log p(\mathbf{y}|i) + \log p(i) - \log p(\mathbf{y}) \\ &\propto \log p(\mathbf{y}|i) + \log p(i). \end{aligned} \tag{5.6}$$

Since the observation for individual factor analyzer is Gaussian distributed:

$p(\mathbf{y}|i) \sim \mathcal{N}(0, \Lambda_i\Lambda_i^T + \Psi_i)$ (i indicates i^{th} FA), therefore Equation 5.6 can be re-written as:

$$\begin{aligned}\log p(i|\mathbf{y}) &\propto \log p(\mathbf{y}|i) + \log p(i) \\ &= -\frac{1}{2}\log|2\pi(\Lambda_i\Lambda_i^T + \Psi_i)| - \frac{1}{2}\mathbf{y}^T(\Lambda_i\Lambda_i^T + \Psi_i)^{-1}\mathbf{y} + \log p(i),\end{aligned} \tag{5.7}$$

where $|\cdot|$ denotes determinant of matrix. Let us define diagonal matrix $\Psi_i = diag(\sigma_{i1}^2, \sigma_{i2}^2, \ldots, \sigma_{id}^2)$. Usually Λ has dimension d-by-k ($k < d$), and $|\Lambda\Lambda^T| = \prod_{j=1}^{k}\lambda_j$ (λ_j is the j^{th} eigenvalue). Therefore we can compute the first part of Equation 5.7 as:

$$\begin{aligned}-\frac{1}{2}\log|2\pi(\Lambda_i\Lambda_i^T + \Psi_i)| = \\ -\frac{1}{2}\left(\sum_{j=1}^{k}\log(\lambda_{ij} + \sigma_{ij}^2) + \sum_{j=k+1}^{d}\log\sigma_{ij}^2\right) + const.\end{aligned} \tag{5.8}$$

In our model, we reduce the k dimensional factor loadings for each analyzer to hold a single column vector ($k = 1$), and it can also be interpreted as, e.g. the 'ray' vector. Furthermore, according to *Woodbury Identity*, the second part of Equation 5.7 can be expressed as:

$$\begin{aligned}-\frac{1}{2}\mathbf{y}^T(\Lambda_i\Lambda_i^T + \Psi_i)^{-1}\mathbf{y} = \\ \frac{1}{2}\mathbf{y}^T(\Psi_i^{-1}\Lambda_i(\Lambda_i^T\Psi_i^{-1}\Lambda_i + I)^{-1}\Lambda_i^T\Psi_i^{-1} - \Psi_i^{-1})\mathbf{y}.\end{aligned} \tag{5.9}$$

Instead of soft assigning data like in *Soft-LOST*, here we *hard* assign data points into m factor analyzers, based on the log posterior probabilities from m FAs. In **M-step**, we re-positions lines to match the points assigned to them. In order words we calculate the covariance matrix of data points within each cluster (FA). As mention earlier we reduce the k dimensional factor loadings to a single column vector, therefore we assign the eigenvector with the largest eigenvalue of each cluster as the new line vector. EM will continue until convergence, and we end up with a mixture of lines $\mathbf{A} = [\Lambda_1, \Lambda_2, \ldots, \Lambda_m]$ to be used as a classifier.

5.1.3 Supervised ICA-like MFA

As introduced earlier, data usually come as pairs in supervised learning. The label set is denoted as $\mathbf{l} = l_1, \ldots, l_T$. As MFA is an unsupervised learning model, we insert manually obtained labels into the ICA-like MFA density model, to

describe the joint distribution of a datum $\mathbf{y}$ and a possible label l:

$$\begin{aligned} p(\mathbf{y}, l) &= \sum_{\mathbf{x}} p(\mathbf{y}, l|\mathbf{x}) p(\mathbf{x}) \\ &= \sum_{\mathbf{x}} p(\mathbf{y}|\mathbf{x}) p(l|\mathbf{x}) p(\mathbf{x}) \\ &= \sum_{i=1}^{m} \sum_{\mathbf{x}} p(\mathbf{y}|\mathbf{x}) p(\mathbf{x}) \quad p(l|i) p(i). \end{aligned} \tag{5.10}$$

The only difference in implementing the unsupervised and supervised ICA-like MFA models, lies in the E-step. We model the log posterior probability $\log p(i|\mathbf{y}, l)$ to add the label information.

$$\begin{aligned} \log p(i|\mathbf{y}, l) &= \log \frac{p(\mathbf{y}, l|i) p(i)}{p(\mathbf{y}, l)} \\ &= \log \frac{p(\mathbf{y}|i) p(l|i) p(i)}{p(\mathbf{y}, l)} \\ &\propto \log p(\mathbf{y}|i) + \log p(l|i) + \log p(i). \end{aligned} \tag{5.11}$$

The comparison between Equation 5.6 and 5.11 shows the difference between log posterior probabilities of the unsupervised and supervised learning models, i.e. the term $\log p(l|i)$.

In the sequel we will compare the performance of these two modified models on various speech related topics. In particular we will train supervised and unsupervised models on the same feature set. For the unsupervised model we first train using only features $\mathbf{y}$. When the density model is optimal, we clamp the mixture density model and train only the cluster tables $p(l|i)$, $i = 1, ..., m$, using training set labels. This is also referred to as **unsupervised-then-supervised** learning. However for supervised learning both feature and label sets are modeled. This is a simple protocol for checking the cognitive consistency: Do we find the same representations when we train them with and without using 'human cognitive labels'?

5.1.4 Illustration of Line Orientations

This pair of models has been applied to various tasks, in which human tagged labels are available. Statistical regularities have been revealed using unsupervised ICA-like MFA model at a variety of chosen time scales, and supervised

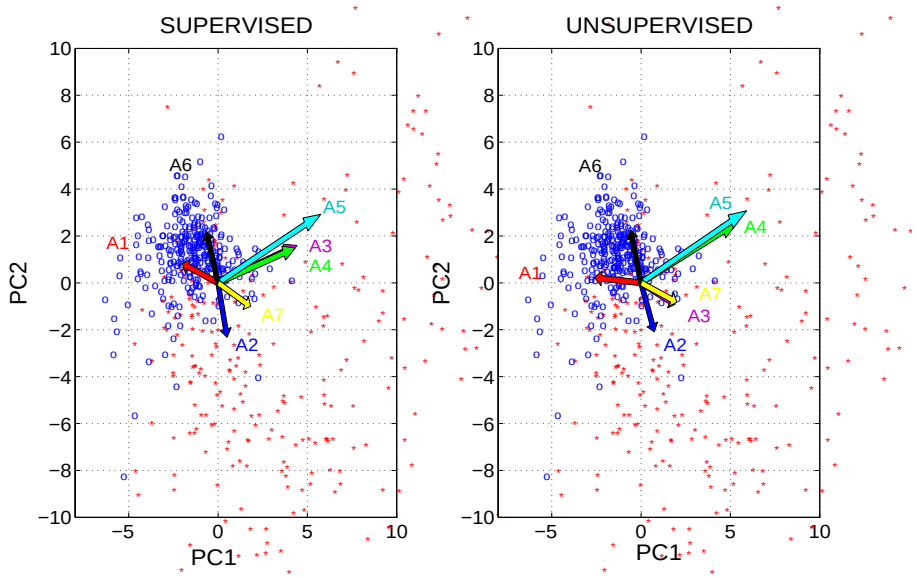

Figure 5.1: Line orientations of ICA-like MFA. It shows the column vectors of matrix Λ of both unsupervised and supervised ICA-like MFA models. Data are denoted by blue 'o' as features extracted from 23 male speakers; and by red '*' as 23 female speakers.

model has been trained with both feature and label sets to predict labels of new input features. Details on experimental design will be discussed in Section 5.3. Here we show the ability of both supervised and unsupervised versions of ICA-like MFA models in discovering line orientations of sparse distributed data. We take one experiment: gender detection based on speech features at time scale 500 $msec$ and 72% remaining energy after EBS. We examine the line orientations, i.e. the column vectors of unsupervised $\mathbf{A}^{unsup}$ matrix and supervised $\mathbf{A}^{sup}$ matrix, and study the degree of similarity. Since it is a binary classification problem, we empirically chose $m = 7$ mixture of FAs. These 7 vectors are shown in Figure 5.1 together with a scatter plot of the training data in the subspace of the first and second principal components. These vectors have been normalized, since the direction rather than the length of the vector is of our interest. Due to the permutation problem of the columns in $\mathbf{A}$ matrices, we aligned columns of $\mathbf{A}^{unsup}$ and $\mathbf{A}^{sup}$ based on the correlation coefficients of the recovered factors $\mathbf{X}^{unsup}$ and $\mathbf{X}^{sup}$. If we hard assign each vector to one class: male or female, A1 and A6 indicate vectors for male speakers, and the rest for female speakers. Orientations of the A_i^{sup} and A_i^{unsup} share high degree of resemblance. As stated in Chapter 3, the distance between vectors is usually

measured by cosine. Input features for both models are 100 dimensions. In this high dimensional space, we computed cosine for each pair of vectors A_i^{sup} and A_i^{unsup}: 1.546, 1.566, 1.481, 0.729, 0.141, 1.576, 0.632. It is not so rational to define a canonical angle between such high dimensional vectors which ranges from 0 to 360 degrees (0 to 2π radians.) However to let the values be more intuitive, we computed the corresponding angles between each vector pairs in a rather low dimensional space, i.e. in the space of 1^{st} and 2^{nd} principal components. The angles are 37.18^0, 13.90^0, 53.38^0, 8.65^0, 0.69^0, 3.97^0, 8.41^0. For some pairs their orientations in the 2D space are almost the same.

5.2 ICA + Bayesian Models

In the previous section, MFA model has been modified to ICA-like density model. As introduced MFA combines the functionality of both FA for dimensionality reduction and GMM for clustering. Each FA is represented by the first eigenvector of the covariance matrix of data points assigned to that particular FA. This modification accommodates spare 'ray-like' data distribution, and share some similarity with ICA model. Here we introduce another set of models, which directly account for the COCA hypotheses: statistical independence and sparse distributed linear mixtures, and include ICA model in unsupervised scheme. Again, having the comparison of the unsupervised and supervised learning in mind, we need to have two models sharing similarities w.r.t the model structure. The Bayesian classifier which assumes a known probabilistic density distribution for each class, has been widely used and is misclassification error rate optimal. Bayesian theory also reflects the *likelihood principle* in perception, and it provides optimal inferences under assumptions. Thus it is capable of revealing plausible perceptual decisions [22]. Here two Bayesian classifiers are chosen. For the unsupervised learning model we first apply unsupervised ICA only on the features. After recovering source signals, we add the label information to a naive Bayes classifier, which assumes that the distribution of the source within each class is Gaussian. To keep the consistency of using Bayesian classifier and Gaussian model, MoG is invoked as the supervised learning model.

Before processing to the devised models, let us take a run-through of naive Bayes classifier and MoG model.

5.2.1 Naive Bayes Classifier

Naive Bayes classifier is a probabilistic classifier. As its name states, it is based on Bayes' theorem:

$$p(\mathbf{C}_i|\mathbf{s}) = \frac{p(\mathbf{s}|\mathbf{C}_i)p(\mathbf{C}_i)}{\sum_i p(\mathbf{s}|\mathbf{C}_i)p(\mathbf{C}_i)} \tag{5.12}$$

where $\mathbf{s}$ is the input data point; $p(\mathbf{C}_i)$ denotes the i^{th} class prior; $p(\mathbf{s}|\mathbf{C}_i)$ is the likelihood of the class $\mathbf{C}_i$; and $p(\mathbf{C}_i|\mathbf{s})$ is the posterior of the i^{th} class given $\mathbf{s} = [s_1, \ldots, s_k]^T$.

Naive Bayes classifier assumes that the effect of a variable value (i.e. s_j) is independent of other values of the variable given the class (i.e. $\mathbf{C}_i$). It is called class conditional independence. This assumption simplifies the computation, and is considered to be 'Naive'. In other words, the dimensions of data are independent, and the likelihood in Equation 5.12 can be rewritten as:

$$p(\mathbf{s}|\mathbf{C}_i) = \prod_{j=1}^{k} p(s_j|\mathbf{C}_i). \tag{5.13}$$

Since $p(s_j|\mathbf{C}_i)$ is learnt from training samples in a given class, and naive Bayes is easy to construct. The conditional independence is a strong assumption, and is rarely true in practice. Nevertheless, naive Bayes is found to perform surprisingly well in classification problems [43]. A biased probability estimation often may not make a difference in classification, since the class with the highest class probability estimate determines the classification, rather than the exact values of probabilities. However for regression problems and probability estimation, naive Bayes shows its deficiency [17, 25].

5.2.2 Mixture of Gaussians

The classic formulation of a linear combination of component densities in a mixture model has the form as:

$$p(\mathbf{y}) = \sum_{j=1}^{n} p(\mathbf{y}|j)p(j). \tag{5.14}$$

where $p(j)$ can be called mixing parameters, or the prior probability of the data point having been generated from component j; and $p(\mathbf{y}|j)$ gives the density of j^{th} component out of n components [9]. This rule has been already used in Equation 5.5 and 5.10.

If the component density is modeled by a Gaussian distribution, a Mixture of Gaussians model is constructed. MoG is one of the most popular models in machine learning. It has shown its superior advantages across various application fields. This method can be tracked back to two decades ago in statistics literature [79]. The model assumes that the data are produced by a mixture of multivariate Gaussians.

MoG is an unsupervised learning technique, it estimates the probability distribution of data with no concern of the label information. Therefore for a classification problem, MoG needs to be applied separately to the data belonging to each class:

$$p(\mathbf{y}|\mathbf{C}_i) = \sum_j p(\mathbf{y}|j, \mathbf{C}_i)p(j|\mathbf{C}_i), \tag{5.15}$$

where $p(\mathbf{y}|j, \mathbf{C}_i) = \mathcal{N}(\mathbf{y}|\mu_{ji}, \Sigma_{ji})$ denotes Gaussian distribution with mean μ_{ji} and covariance Σ_{ji}, $p(j|\mathbf{C}_i)$ is the mixing parameters in class $\mathbf{C}_i$. Parameters μ, Σ are estimated from training sets via the standard Expectation-Maximization (EM) algorithm.

For prediction, posterior probabilities of MoG models from each class are calculated, according to:

$$p(\mathbf{C}_i|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{C}_i)p(\mathbf{C}_i)}{\sum_i p(\mathbf{y}|\mathbf{C}_i)p(\mathbf{C}_i)}, \tag{5.16}$$

Observations are assigned to the class having the maximum *posterior* probability. Maximum A *Posteriori* (MAP) criterion aims at maximizing the *posterior* $p(\mathbf{C}|\mathbf{y})$ rather than maximizing the likelihood $p(\mathbf{y}|\mathbf{C})$.

5.2.3 Unsupervised Model

The particular unsupervised model in COCA analysis obeys the **unsupervised-then-supervised** scheme. Independent sparse sources will be first recovered by ICA. To unveil the capability of ICA in class label prediction, an supervised Bayesian model: naive Bayse, will be followed.

As we know typical algorithms for ICA use centering, whitening and dimensionality reduction as preprocessing steps in order to reduce the complexity of the algorithm. PCA is normally used to achieve these steps. Since in the preprocessing pipeline we have applied PCA on stacked and sparsified MFCC features, we directly apply ICA algorithm to PCA coefficients.

The component y_i of the observation vector $\mathbf{y} = (y_1, \ldots, y_d)^T$ is generated by summing independent sources $\mathbf{s} = (s_1, \ldots, s_k)^T$ with different mixing weights

$a_{i,j}$:

$$y_i = a_{i,1}s_1 + \ldots + a_{i,j}s_j + \ldots + a_{i,k}s_k. \tag{5.17}$$

The noise free ICA model is a simplification of the hidden variable model, as we have encountered several times. What worth to emphasize here is that the k independent sources $\mathbf{s}$ are assumed non-Gaussian. ICA aims at estimating both the mixing matrix Λ and sources $\mathbf{s}$. This is done by either maximizing the non-Gaussianity of the calculated sources or minimizing the mutual information. If Λ is a square matrix, original sources can be recovered by

$$\mathbf{s} = \mathbf{W}\mathbf{y}, \tag{5.18}$$

where $\mathbf{W} = \Lambda^{-1}$ is the unmixing matrix.

To reveal the performance of unsupervised learning in classification tasks, we first train the unsupervised model using only features $\mathbf{Y}$ to recover the sources $\mathbf{S}$. Since sources are independent, then naive Bayes classifier can be applied to sources with the training set labels. Based on Equation 5.13, we model the class conditional probability of each independent variable $p(s_j|\mathbf{C}_i)$ as univariate Gaussian distributed $\mathcal{N}(\mu_{ji}, \sigma_{ji}^2)$. For the classification problem, in the training phase, we will learn $\mathbf{W}$ from training data, and recover the sources $\mathbf{S}^{train}$. They are, in turn, input to naive Bayes classifier to learn model parameters $\{\mu_{ji}, \Sigma_{ji}\}$. To predict the label for a new datum $\mathbf{y}^{new}$, $\mathbf{W}$ learnt from training data will help in Equation 5.18 to recover $\mathbf{s}^{new}$. Whereafter, the trained naive Bayes classifier with a set of Gaussian parameters will be used on $\mathbf{s}^{new}$ to calculate the posterior probability for each class, based on Equation 5.12. The class with *MAP* indicates the predicted class label of the new datum.

5.2.4 Supervised Model

For the supervised learning model, we intend to choose one of the flexible models. Bayesian classifiers are concerned as misclassification error rate optimal, hence we use MoG to model variate cognitive classification problems.

As MoG is potentially an unsupervised model, we need to separate training data based on their labels into C classes, and build a MoG model on each class. For simplicity, we assume the covariance matrices Σ_{ji} to be diagonal. Thus axes of the resulting Gaussian clusters are parallel to the axes of the input data space. Note that although features are independent within each mixture component due to the diagonal covariance matrix, the mixture model does not factorize over features. The MoG is capable of modeling arbitrary dependency structures among features [9], if the number of mixture components is sufficiently large. On

the other hand, a MoG with many mixture components is prone to overfitting, and will most likely not generalize well. In our experiments, we vary the number of mixture components, and select models according to classification accuracy. New data are assigned to the class having the maximum *posterior* probability.

5.3 Experimental Design

5.3.1 Database Description

Two pairs of COCA models to loosely represent statistical regularities/independence and human cognition, are utilized to data gathering from TIMIT database [27]. TIMIT collects reading speech from 630 native American English speakers from 8 major dialect regions. Each speaker reads 10 sentences in total, and each sentence lasts approximately 3 *sec*. The database has been suggested into training and test sets, where 168 out of 630 speakers were allocated into test set. For each utterance, three associated transcription are provided: Associated orthographic transcription of the words; Time-aligned word transcription; Time-aligned phonetic transcription. In the meanwhile, the speaker information, e.g. gender, age, height, race and education were also recorded. Hence we could obtain several labels that we think as cognitive indicators, which are labels that humans can infer given sufficient amount of data.

Phonemic and phonetic symbols used in TIMIT lexicon are given, and they are divided into refined categories, including stops, affricates, fricatives, nasals, semivowels and glides, vowels and others. Stops and affricates are further given the closure symbols: e.g. 'dcl' for stops 'd', and 'tch' for affricates 'ch'. Others in TIMIT definition includes different types of silence, pause and non-speech segments. Totally, 64 symbols are used in transcription. The height of all speakers locates in the range from $4'9''$ to $6'8''$ with 22 values. The age of the TIMIT speakers is not evenly distributed either: around 60% speakers are within 21 to 30 years old; and about 30% within age 60 to 72. Figure 5.2 shows the histogram of the height and age information of 630 speakers.

5.3.2 Data Set Construction

In Chapter 4 we have proven that phoneme and speaker identity as cognitive components of speech. We envision that gender, age and height are potential cognitive indicators as well. We have carefully selected a sufficient amount of data to reach the computational limits of the PC (Intel Pentium IV computer

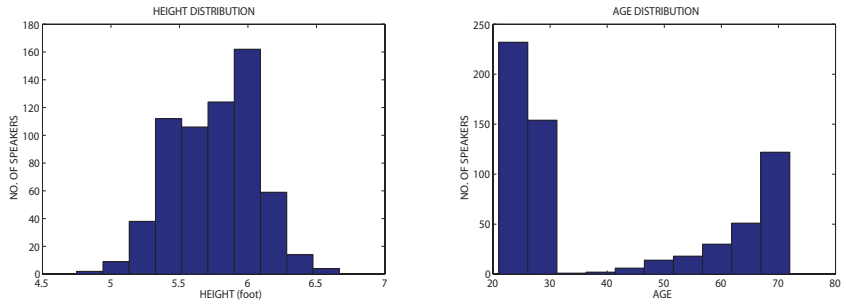

Figure 5.2: Histogram of height and age distribution of TIMIT speakers.

with $3GHz$ and $2GB$ of RAM), in the meanwhile we have guaranteed that the data represent the breadth of available information in the database. We chose 46 speakers with equal gender partition, and speech signals cover 60 relatively important phonemic and phonetic symbols, including all the phonemes. To simplify the classification problem, we pre-grouped phonemes into 3 large categories: vowels, fricatives and the rest. As to height, the chosen speakers cover all the 22 height values in the TIMIT database, and to keep the distribution even within each group, we pre-grouped speakers into 6 classes: height from $4'9''$ to $4'10''$; $5'$ to $5'2''$; $5'3''$ to $5'6''$; $5'7''$ to $5'10''$; $5'11''$ to $6'2''$; and $6'3''$ to $6'8''$. The age of the chosen speakers covers the full range: 21 to 72. Same as before, we pre-grouped ages into 4 sets, in order to keep an approximate even population among sets: from age 21 to 25; 26 to 29; 30 to 59; and 60 to 72, both endpoints are included in the set. Also we speculate that with sufficient amount of data, humans are able to guess speakers height and age within a range from their speech.

5.3.3 Experiment Construction

Since cognitive components of phonemes and identities were found at different scales, to discover higher cognitive components, we studied features at different time scales. The unsupervised and supervised models were compared in a set of experiments: we stacked the basic time scale features into several longer time scales, and sparsified the stacked features with different degrees. We anticipated to find out the role of the time scale, in the meanwhile to examine the role of sparsification. In a particular condition (a certain time scale and sparsification level), the same feature sets have been used in the above mentioned five classification tasks for both unsupervised and supervised learning models, and the difference among them was the manually obtained class labels.

Following the preprocessing pipeline, we first extracted 25-dimensional MFCCs from speech signals. The 0^{th} order MFCC, which represents the total log energy of each short-time frame, was also included. To study the role of time scale, we stacked the basic features into a variety of time scales, from basic time scale up to above $1s$ (20, 60, 100, 150, 300, 500, 700, 900 and $1100ms$). The degree of sparsification was controlled by thresholds. The sparsification was carried out on the normalized stacked MFCCs. PCA was applied on stacked and sparsified features, and dimensionality of features was reduced. For features at longer time scales than 20 $msec$, their dimensions were reduced to 100, and the dimension of the features at the basic time scale remained the same, i.e. 25.

The signals from the first 6 sentences of each of 46 speakers, were used as the training set, and were processed following the preprocessing pipeline. The outcomes were input into the unsupervised and supervised models respectively. The number of mixtures for MFA models were set empirically, basically we chose more mixtures than the actual number of classes. As to MoG, for each experiment with certain time scale and sparsification degree, we built models with different number of mixtures (n). The model selection determined the final n value. Appendix A lists out the final number of mixtures for each experiment. For prediction we preprocessed the test set, which consisted of the rest 4 sentences of the 46 speakers, following the same procedure. For probability models, we could access to both predicted labels and posterior probabilities.

5.4 Comparison Methods

From the unsupervised learning model and supervised learning model, we get access to classification results. The comparison is defined at three levels. In appendix E, we compared the results in phoneme, gender, identity and height classifications in a straightforward way by looking at error rates, and the error rate correlation. To examine the accuracy of a classifier, error rates are sufficient. However they seem a bit superficial to measure the likeness of two models, since the discrimination of individual samples has been covered up by the average. More detailed comparison has been utilized in appendix F, G and H. Sample-to-sample based comparison reveals the confusion of both models. A sample based posterior probability comparison gives us the intuition on how certain or uncertain classifiers are while making a specific decision.

Table 5.1: Recommended time scales for modeling Phonemes, Gender, Age, Height, Identity.

(ms)	**Phoneme**	**Gender**	**Age**	**Height**	**ID**
Timescale	20	300-500	500 < t <1000	$\geq$ 1000	> 1000

5.4.1 Error Rate Comparison

Representations of unsupervised and supervised learning on both training and test sets have been investigated. Here let us first focus on classification error rates. A series of experiments at different time scales and with different degree of sparsification have been carried out. The error rates as a function of time scale can be represented as a curve in a time scale vs. error rates plot. The sparsification degrees give the figure with different curves. Figure 5 in appendix E is one example in gender detection using ICA-like MFA models; and Figure 3 in appendix F is another example in phoneme classification using ICA + Bayesian models. By investigating the tendency of curves, we could find out the approximate time scale, at which the cognitive task was best modeled. The tendency of unsupervised and supervised models did agree, and the estimated time scales from ICA+Bayesian models coincide with those from ICA-like MFA models in the same classification tasks. The recommended time scales given by our COCA models are summarized in Table 5.1. Time scales are coarsely given in ranges, and may fluctuate depending on models.

Another point revealed by these figures is that if we do cross-model comparison on both training and test sets, the similarity between unsupervised and supervised learning is obvious. To have a close look at the recognition error rates, we measured the correlation of test error rates, and displayed them in unsupervised vs. supervised form, see Figure 4 in appendix G as an instance. High correlation between error rates of the paired models indicate similarity of the representations. The correlations of all tasks were distinguished, where identity recognition deserves our special attention: for the given time scales and thresholds, data located closely along $y = x$, with correlation coefficient $\rho = 0.9660$, and $p < 4.04 \times 10^{-38}$.

5.4.2 Sample-to-Sample Based Comparison

While error rates show the overall classification performance, the error of each sample tells the pattern of making decision, such as the temporal locations where it is most likely to have wrong predictions, and whether the wrong predictions

come in a row or randomly, further whether two models make mistakes or correct decisions the same way or at the same place. Therefore we form a sample-to-sample based comparison. On this basis, two approaches are proposed. The first one studies the correlation of two models in decision making by investigating the matching or mismatching degree in percentage. The second one is more intuitive. We present the classification results of both models, and study the decision making pattern with respect to the ground truth to show where these two models make the same decisions.

First we computed both correctly classified sample rate by unsupervised and supervised models for the test set of a given task r_{cc}, both wrongly classified sample rate r_{uu}, and the disagreement of two models: correctly classified by supervised model, but wrongly classified by unsupervised model r_{cu}, vise versa i.e. r_{uc}. The total error rates of both models are defined as r_{sup} standing for supervised model; and r_{usup} for unsupervised model. To eliminate the bias caused by total error rates of each model, we thus introduced a new set of rates:

$$\begin{aligned} R_{cc} &= \frac{r_{cc}}{(1-r_{sup})(1-r_{usup})}, \quad R_{uu} = \frac{r_{uu}}{r_{sup}r_{usup}}, \\ R_{cu} &= \frac{r_{cu}}{(1-r_{sup})r_{usup}}, \quad R_{uc} = \frac{r_{uc}}{r_{sup}(1-r_{usup})}. \end{aligned} \tag{5.19}$$

The first row in Equation 5.19 gives the rates for the matching case; whereas the second row shows the mismatching rates. Finally to keep the rates as percentage, we normalized them by their summation:

$$P_{ij} = \frac{R_{ij}}{\sum_{mn}(R_{mn})}, \quad m,n = (c,u); \quad i,j = (c,u). \tag{5.20}$$

Appendix H includes the results of sample-to-sample error comparison in four tasks: phoneme classification, gender detection, age detection and speaker identification. The comparison was between unsupervised ICA+naive Bayse model and supervised MoG model on the test set. In the subplot, the lower left circle refers to the normalized both correctly classified sample rate by unsupervised and supervised learning: P_{cc}; upper right one stands for P_{uu}. The diagonal circles show the disagreement of two schemes in making decisions: P_{cu} upper left; P_{uc} lower right. The area of each circle represents the portion in percentage, and the four areas sum to 1. The plot reveals that to what degree representations derived from supervised and unsupervised learning match, and how well they match with human labels (the ground truth). A large percentage allocated on the off-diagonal circles, indicating high correlation between supervised and

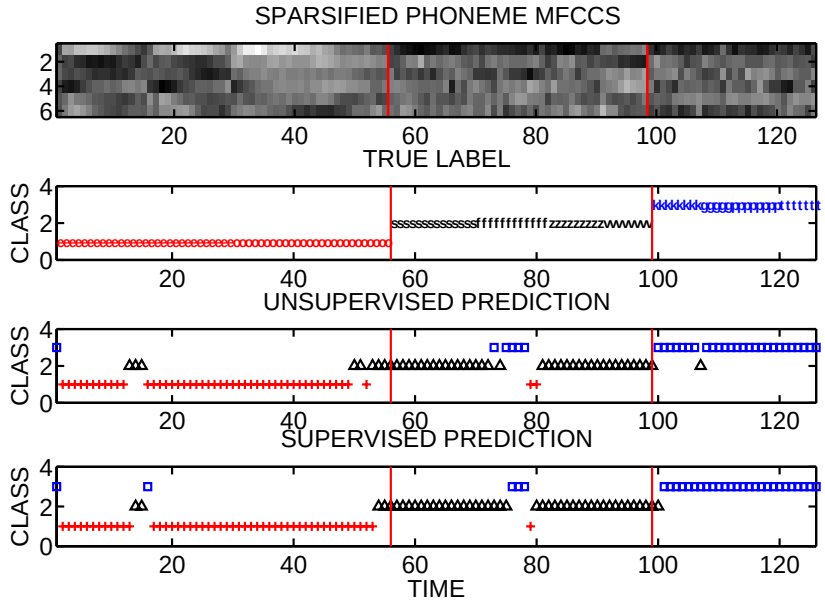

Figure 5.3: Sample-to-sample phoneme classification among vowels, fricatives and stops. The first panel shows the temporal development of MFCCs. Boundaries of 3 phoneme classes are highlighted by vertical lines. The second panel gives the true labels, denoted by phonetic symbols. The last two panels give the unsupervised and supervised label predictions, marked by 3 shapes. The decision patterns between supervised and unsupervised learning show high similarities.

unsupervised learning. The overall comparison under all conditions was summarized as a histogram shown on the left-hand side of the figures. The locations of bars correspond to the circles in other subplots.

To get a direct observation of the classification performance, we study the predicted class labels instead of error rates. One example given here was carried out on three groups of phonemes: vowels eh, ow; fricatives s, z, f, v; and stops k, g, p, t, where eh stands for the vowel in the word 'BET', and ow for the vowel in 'BOAT'. Figure 5.3 presents the sample-to-sample classification results of unsupervised ICA + naive Bayes and supervised MoG model. MFCC features were first sparsified with 99% remaining energy, and then PCA reduced the dimension to 6, and the resulting features were modeled by unsupervised and supervised learning methods separately. The pattern of prediction of both models are alike, and the percentage of matching (correct predictions from both models and misclassified samples from both models) between supervised and unsupervised learning was up to 91%. Two models tend to make mistakes in similar areas. In all places that supervised learning made wrong predictions,

unsupervised learning model also had wrong predictions. Moreover, error prediction in most cases does not show up alone, meaning features representing some particular phonetic symbols are easier to be mistaken than others.

5.4.3 Posterior Probability Comparison

So far we have seen that the unsupervised and supervised learning models bear close correspondence at the level of error rates and sample-to-sample classification. A more detailed comparison can be obtained by considering posterior probabilities on a sample-to-sample base. In the above mentioned matching case, i.e. both models make the same predictions either correct or wrong, we can measure the certainty of these decisions, and compare them pair to pair between unsupervised and supervised models. The comparison again has been carried out on four cognitive tasks involving phonemes, gender, age and identity. Within each task, one or more model from a certain experiment has been illustrated in appendix H. They are female model in gender detection; fricatives model in phoneme classification; 12 speaker models in identity recognition; and 21-25 age model in age detection. The data shown in the figures belong to the corresponding set, meaning the truth labels for data are 1 using 1-of-C coding. If two models are the exact match, we expect that the posterior probabilities locate along the diagonal of the histograms with high distribution at $(1, 1)$ in the coordinate system, which corresponds to the correct decisions by both models, and at $(0, 0)$ referring to the wrong decisions by two models. High percentage falls into $(1, 1)$ and $(0, 0)$ in the coordinate system, indicating that unsupervised and supervised models do match to a certain degree.

5.5 Summary

We introduced the high-level cognitive component analysis in this chapter. We devised two pairs of unsupervised and supervised models. Both model sets are able to accommodate sparse 'ray structure', and within each set two type of learning models share similarities w.r.t the model structure. The first pair of models employed the transformation ideas from the *LOST* models, and converted MFA model into a ICA-like density model. EM algorithms on top of hard assigning data points into line vectors, were applied to determine the line orientation vectors. Similar to unsupervised ICA-like MFA model, we modeled the joint probability of features and labels in supervised version. Appendix E provides us with some experimental evidences that unsupervised and supervised learning did obtain similar representations. The line vectors defined by

both learning models were shown in 2D space for illustration in subsection 5.1.4.

Even though the first pair conveyed the statistical independence by reducing the k dimensional factor loading matrix of each FA into a single column vector, the independency was not revealed explicitly. Subsequently, ICA+naive Bayes model has been combined as unsupervised learning model to be compared with supervised MoG model. The naive Bayes classifier was applied to the recovered independent sources by ICA, to present the classification capability of the unsupervised learning. MoG models were assumed with diagonal covariance matrices, which made the Gaussian clusters parallel to the axes of the feature space.

The experiments with two pairs of models were described. These basic setups were utilized in all the publications on the high-level COCA. For detailed comparison between unsupervised learning to reveal statistical regularities and supervised learning of human labels, we carried out the cross-model contrast at three levels, from macro-scale to micro-scale. The classification performance is normally represented by error rates. Thus we first investigated the time scale vs. error rate results. Intuitively, the tendency of the curves did agree to some extend, indicating the similarity between unsupervised and supervised learning. These cognitive tasks were best modeled at different time scales, indicating different levels of cognitive functions. Further we compared unsupervised learning vs. supervised learning w.r.t error rates. To delve into the performance, we studied the errors on the sample-to-sample basis. It was done in two ways: we measured the matching and mismatching degrees in percentage; and we looked at the prediction results sample by sample in a small data set aiming at looking for the decision making patterns by two learning methods. High percentage was found in the matching case for all the tasks, and the prediction patterns were quite alike: first when supervised learning made mistakes, it is very certain that unsupervised learning had the same mistakes; and wrong predictions very often came together, indicating these samples share certain characteristics rather than some random samples. Finally to test the certainty of a sample-based decision, we further dig into the posterior probabilities provided by two learning methods when they agreed in their decisions. The results also fell into our expectations. The certainties of a single decision matched for both models, i.e. when one model was certain, the other was certain as well, and vice versa. These were shown by having much bigger values in the histogram of posterior probabilities comparison at (1,1) and (0,0) in the coordinate system.

Chapter 6

Conclusion

Cognitive Component Analysis was proposed as a way to investigate the consistency of statistical regularities in a signaling ecology and human cognitive activity. An unsupervised learning algorithm is defined as cognitive component analysis if the ensuing group structure is well-aligned with that resulting from human cognitive activity. The hypothesis of COCA stands on two bases: statistical independence and sparse representation. It is basically ecological: we envision that features which is essentially independent in a context defined ensemble, can be efficiently coded as sparse independent component representations. The two main points of COCA are the unsupervised grouping of data, so as to extract statistical regularities; and the comparison between the representations derived from human cognition and the regularities extracted from perceptual inputs.

Since COCA works with perceptual inputs, especially speech as the form of sound waves, the actual data for modeling should be able to represent the information processed by the human auditory system. The anatomy or physiology of the auditory system can explain some aspects of auditory perception, while psychophysical experiments and perceptual studies also help us to understand the perception. Thus we opened this thesis with a brief description of human cognition with emphasis on the physiology of human auditory system. Based on our understanding of auditory perception, the classic preprocessing pipeline was built to prepare raw waveforms into representative features. The design

of the pipeline took two aspects into account. The choice and sequence of the techniques tried to emulate functions of peripheral auditory system and mimicry psychoacoustics. Secondly, we tried to use standard speech signal processing techniques. This pipeline consists of feature extraction, feature integration, sparsification and principal component analysis. The chosen feature: mel-frequency cepstral coefficient follows the logarithmic dependence of the signal power, in the meanwhile it maps the linear frequency into mel-scale using critical band filters, and both points are in line with the human auditory system. MFCC loosely represents the human auditory system, especially the best understood peripheral auditory system, but it fails to contain the information for sound localization and loudness accuracy, which are mainly determined by the folds of pinna in the outer ear. Nevertheless, sound localization and loudness accuracy are not the essential information interesting to COCA analysis. Feature stacking as the simplest form for feature integration aims at the temporal integration of short-time features into longer time scales, since different cognitive tasks may be best modeled at different time scales. Energy based sparsification filters out small signals, which attempts to mimicry the pattern of cortical neurons firing rates, and it saves energy. Finally principal component analysis, or latent semantic indexing in textual study, is regarded as the basis for all cognitive processing, and it is claimed to have human-like performance.

This dissertation has mainly concentrated on speech signal processing, even though COCA is a generic tool which has been applied to diverse topics for discovering cognitive related components. A number of examples were introduced in Chapter 3 on textual analysis to reveal latent semantics, on music data to look for genre-specific structure, and on social network to locate communities of actors. All these instances involved unsupervised learning method (e.g. PCA or ICA) aiming at statistical regularities.

Except for the introduction and foreshadowing, this dissertation was divided into two parts. The sequence of this report followed the development of COCA analysis of speech data. We unveiled COCA study from the low-level cognitive components. The promising findings impelled us to generalize them to higher level cognitive functions. We devised a protocol for COCA to measure the correlation between statistical properties (especially independence) and human cognition.

6.1 The Low-level COCA

On low-level COCA, we allocated the effort to phoneme recognition and speaker identification based on speech. Unsupervised learning has early been proven to

be able to discover statistical regularities. In this study, we followed the COCA preprocessing pipeline, and focused on the resulting 'ray structure' in the feature space, which was derived by unsupervised learning methods: e.g. PCA and ICA.

Unsupervised learning based on sparse linear component analysis of speech signals discovered cognitive relevant components, representing the universal linguistic atoms namely phonemes and speaker-specific features in appendix B. These findings impelled us to ponder on whether humans use such theoretical optimal representations in other abstract and higher level perceptual tasks as well.

In phoneme study, we speculated that we have found the 'invariant cue' in perceived signals. The characteristics of speech signals vary from speaker to speaker, and even from trial to trial within one speaker, which may be caused by different health conditions and emotions. However the perceived signals are often some stable phonetic features, in which the key features follow an invariant form. This is the basic concept for the acoustic 'invariant cue'. Except for the findings of /e/ sound opening both letter 's' and 'f', we delved into the phoneme study with an attempt to search for quantitative evidences for 'invariance cue' in various conditions. The common phonetic unit shared by two letters from one speaker, has been found locating along 'rays' in the phonetic space derived by PCA. The same phenomenon has been revealed in the multi-speaker case, with 3 speakers pronouncing sound /ti:/. The 'invariant cue' by the definition as the invariants derived from phoneme units did exist, and could be found by simple sparse linear component analysis. Applying ICA after PCA, we are not restricted to orthogonal basis vectors, thus the resulting receptive fields have overlaps, which in turn can detect more subtle differences than 'orthogonal' receptive fields. ICA, given the right representation, has been proven as a generic tool for COCA. Examples are given in appendix C and D.

In speaker identity study, speaker-specific cognitive components were found at a longer time scale (1 *sec*) by using PCA as well. A intriguing result was obtained while using the same text for all speakers. Two phenomena were revealed together. First of all, phoneme-like structures did exist since the same text grouped together in the semantic space. Speaker-specific structures has also been found, due to the evidence that features from one individual did spread along similar orientations with offsets in the space. We speculated that these phenomena were the results of the interaction between the phoneme-like effect and speakers' 'voiceprint'.

6.2 Unsupervised Learning vs. Supervised Learning

To answer the question whether humans use theoretical optimal ICA representations in higher level perceptual tasks, we moved on to the higher cognitive functions, and devised a protocol to test the consistency of statistical regularities of inputs and human cognitive activity. Unsupervised learning can discover statistical regularities. However human cognition is complex and sophisticated, and not yet fully discovered and understood. One information which is much easier to access and model, is the human behavior, and it is seen as the direct consequence of human cognition. Thus we represented cognition as a classification rule in supervised learning of manually obtained labels. This interpretation is not comprehensive, however it is capable of representing some intrinsic mechanism of human cognition. COCA is not limited to one specific technique, but rather a conglomerate of different techniques. For comparison purpose, we designed two pairs of models. Both pairs of models were modified or designed to accommodate the data representations, namely sparse 'ray structure'. To have a fair comparison, the unsupervised learning and supervised learning models shared certain similarities in model structure.

In appendix E, we introduced the proposed ICA-like MFA density models. MFA was modified as a mixture of line orientation vectors, and factor loadings for each FA was reduced to a single column vector. Thus the modified MFA with k mixtures have k line vectors. The unsupervised-then-supervised scheme was used to show the classification results. The supervised version of ICA-like MFA modeled features together with the corresponding labels. The independency was not reflected quite obviously in the first pairs of model. To convey the statistical independence, an unsupervised learning model with ICA + naive Bayes model has been constructed, to compare with Mixture of Gaussians, and they are covered in appendix F, G and H. Following the same unsupervised-then-supervised scheme, ICA first recovered independent sources, then naive Bayes classifier was used on sources to perform the classification. Whereas MoG was used directly on the mixed data.

Having two pairs of models in hand, we did cross-model comparison within each model set. The unsupervised learning intended to find statistical independence, and the supervised learning attempted to roughly represent human cognition. The contrasts between unsupervised and supervised learning have been carried out systematically. It was divided into three levels: classification error rates, sample-to-sample errors, and the sample based posterior probabilities. Inspired by previous findings that phoneme cognitive component was found at a short time scale, whereas speaker identity was best modeled at a longer time scale,

i.e. 1 *sec*, we modeled features at different time scales to look for high-level cognitive components. Various cognitive tasks on, such as gender, height, age, phoneme and identity, were found to be best modeled at various time scales. Further the error comparison at the sample-to-sample level showed high percentage of matching in making the same decisions, either correct or wrong decisions, from unsupervised and supervised learning. The prediction patterns from both learning methods have been illustrated, and the representations resembled each other. The posterior probabilities comparison measured the certainty of one predication, and it was often the case that when one model was quite certain about a decision, the other one also had the probability on the same level.

In summary, the preliminary study of COCA indicated that statistical regularities can be revealed by simple sparse linear component analysis, and ICA applied after the preprocessing pipeline can relax the orthogonal basis, and allow the model to detect more subtle differences among features than 'orthogonal' receptive fields. The consistency between statistical regularities/independence and human cognition can be tested using the devised protocol: unsupervised learning vs. supervised learning. A detailed contrast scheme from classification error rates, sample-to-sample error, to posterior probability level, measures the matching or mismatching degree of two learning methods, representing statistical independence and human cognition. Indeed, the two classifications agree on a majority of scenarios in several cognitive tasks related to speech perception, from low-level to high-level. Age and height were also studied from speech signals, corresponding to the human ability to guess speakers age and size in a rough range. All in all, the unsupervised learning algorithm and supervised learning proxy for a human cognitive activity did lead to comparable classifiers.

6.3 Future work

In the period of this project, some ideas have become clear and appealing, and we believe that they have a great potential to lead the work of COCA to a new research direction.

Onset dynamics is an important aspect for speech perception, and it has caught more and more attentions lately. In the psychophysics standpoint, onsets on the basilar membrane is of interests. Even though MFCC loosely reflects the human auditory system, and represents the log energy distribution over the basilar membrane, it gives the same weight on onsets and offsets by means of providing feature vectors on the same level. However it has been proven that onsets are more crucial than offsets in speech production and perception. Therefore to discover new features to emphasize the significance of onsets may

be a possible direction for COCA.

To obtain features at longer time scales than the basic one, e.g. 20 *msec*, simple feature stacking has been used. However stacking only models the temporal development in a very loose way, and the resulting features after principal component analysis only keep the important information with respect to the variance among those stacked short-time features. The multivariate autoregressive model, on the other hand, is capable of modeling the temporal dynamics and the dependency of feature dimensions as well. We think MAR might be a good substitute for feature stacking. The order of AR can be looked into, in order to get the right representations.

Appendix A

The Number of Mixtures

The number of mixtures for both ICA-like MFA models and ICA + Bayesian Models are listed out. For ICA-like MFA models, we used the same number of mixtures within each cognitive task. That is for a certain task, at any time scale and any sparsification degree, we used a fixed number of ICA-like MFA models for both unsupervised and supervised models. Table A.1 gives the number of mixtures together with the number of classes for phoneme classification, gender detection, height estimation and speaker identification.

We systematically did model selection for MoG models in ICA + Bayesian Models set. For comparison purpose, we chose the best model in each condition: at a certain time scale and sparsification degree. Tabel A.2 to A.5 gives the number of mixtures for all the models used for performance comparison in phoneme classification, gender detection, age detection and identity recognition.

Table A.1: Number of mixtures for ICA-like MFA models.

	phoneme	gender	height	identity
No. of classes	3	2	6	46
No. of mixtures	9	7	20	80

Table A.2: Number of mixtures for phoneme MoG models.

	100%	99%	97%	90%	85%	75%	72%	65%
$20ms$	40	40	30	30	20	30	30	30
$100ms$	20	40	20	20	40	10	30	40
$150ms$	20	25	15	15	25	25	15	25
$300ms$	10	15	12	10	5	5	12	15
$500ms$	2	6	4	8	6	4	10	8
$700ms$	4	4	6	6	6	5	3	3
$900ms$	4	3	2	4	5	5	3	3
$1100ms$	3	3	3	4	4	3	4	4

Table A.3: Number of mixtures for gender MoG models.

	100%	99%	97%	90%	85%	75%	72%	65%
$20ms$	280	190	190	100	40	10	10	10
$100ms$	70	100	130	70	130	160	160	100
$150ms$	70	100	130	40	70	40	10	5
$300ms$	70	40	5	160	70	100	5	10
$500ms$	5	30	10	10	30	30	10	50
$700ms$	5	5	5	70	5	10	5	5
$900ms$	90	50	50	10	30	110	5	30
$1100ms$	30	10	10	30	10	30	70	70

Table A.4: Number of mixtures for age MoG models.

	100%	99%	97%	90%	85%	75%	72%	65%
$20ms$	120	100	120	120	60	40	120	120
$100ms$	120	40	16	30	60	16	58	30
$150ms$	20	42	40	40	120	80	40	22
$300ms$	32	20	8	14	32	32	8	8
$500ms$	40	11	15	19	20	27	100	100
$700ms$	6	10	18	20	14	22	18	2
$900ms$	8	8	80	20	14	40	20	120
$1100ms$	5	5	8	120	120	8	8	80

Table A.5: Number of mixtures for identity MoG models.

	100%	99%	97%	90%	85%	75%	72%	65%
$20ms$	40	20	20	15	25	15	5	70
$100ms$	20	40	100	10	15	20	5	15
$150ms$	8	2	2	5	5	2	5	5
$300ms$	2	2	2	2	2	4	2	2
$500ms$	2	2	2	2	3	2	2	2
$700ms$	2	2	2	2	3	2	3	2
$900ms$	2	2	2	2	2	2	2	2
$1100ms$	2	2	2	2	2	2	2	2

Appendix B

On Low-level Cognitive Componnet Analysis

This article is published in *Proc. International Conference on Computational Intelligence for Modelling* 2005, vol. 2, pp 852-857, with the same title. Authors are Ling Feng and Lars Kai Hansen. It is also available as IMM publication database with number imm3664.

On Low-level Cognitive Components of Speech

Ling Feng
Intelligent Signal Processing,
Informatics and Mathematical Modeling,
Technical University of Denmark, Denmark
lf@imm.dtu.dk

Lars Kai Hansen
Intelligent Signal Processing,
Informatics and Mathematical Modeling,
Technical University of Denmark, Denmark
lkh@imm.dtu.dk

Abstract

In this paper we analyze speech for low-level cognitive features using linear component analysis. We demonstrate generalizable component 'fingerprints' stemming from both phonemes and speakers. Phonemes are fingerprints found at the basic analysis window time scale (20 msec), while speaker 'voiceprints' are found at time scales around 1000 msec. The analysis is based on homomorphic filtering features and energy based sparsification.

1. Introduction

The human perceptional system can model complex multi-agent scenery. It is well documented that humans use a broad spectrum of cues for analyzing perceptual input and for identification of individual signal producing agents, such as speakers, gestures, affections etc. Such unsupervised signal separation has also been achieved in computers using a variety of independent component analysis (ICA) algorithms [1]. It is an intriguing fact that representations found in human and animal perceptual systems closely resembles the theoretically optimal representations obtained by independent component analysis on visual contrast detection [2], on visual features involved in color and stereo processing [3], and on representations of sound features [4].

Ref. [5] defined and investigated the independent *cognitive* component hypothesis, which basically asks the question: *Do humans also use these information theoretically optimal 'ICA' methods in more generic and abstract data analysis.* We proposed to use the term *cognitive component analysis* (COCA) for unsupervised learning algorithms that present such 'spontaneous cognition'.

Here we are interested in pursuing this idea in the context of speech. We are interested in purely auditory aspects, not contents *per se*. We will focus on two aspects, phoneme features and speaker features. Our presentation will be qualitative, mainly based on simple visualizations of data, thus we avoid unnecessary algebraic complication.

Grouping of events or objects in more or less distinct categories is fundamental to human cognition. In machine learning, classification is a rather well-understood task when based on *labeled* examples [6]. In this case classification belongs to the class of *supervised* learning problems. On the other hand clustering which is related to *unsupervised* learning problem, uses general statistical rules to group objects, without a priori providing a set of labeled examples. It is a fascinating finding in many real world data sets that the label structure discovered by unsupervised learning closely coincides with labels obtained by letting a human or a group of humans perform classification, labels derived from human cognition. Grouping by ICA has been earlier pursued for several abstract data types including text, dynamic text (chat), images, and combinations hereof, see e.g., [7, 8, 9, 10, 11]. It was found in these research works that ICA is a more appropriate model than both principal component analysis (PCA), which is too constrained, and clustering, which may in some instances be too flexible as a representation of text data [5].

2. Cognitive component analysis

Lee and Seung introduced the method of non-negative matrix factorization (NMF) [12] as a scheme for parts-based object recognition. The factorization of an observation matrix in terms of a relatively small set of *cognitive components* leads to a parts-based object representation. The values of the non-negative representation for objects in images and text have been demonstrated. In 2002, similar parts-based decompositions were obtained in a latent variable model based on non-negative linear mixtures of non-negative *independent* source signals [13]. Holistic, but

parts-based, recognition of objects is frequently reported in perception studies across multiple modalities and increasingly in abstract data, where object recognition is a cognitive process. Together these findings are often referred to as instances of the more general *Gestalt laws*.

2.1. Latent semantic indexing (LSI)

Principal component analysis (PCA) is a very useful tool for dimensionality reduction and may be used to find group structure in data when the signal-to-noise ratio is high. PCA has been used for basic perceptual feature analysis, such as in images under the name Karhunen-Loeve transform [14], and for analysis of abstract data such as text under the name latent semantic indexing (LSI) [15]. Our approach is inspired by LSI, and the main innovation here is the active search for generalizable non-orthogonal linear features that may be described in terms of an independent component generative model.

Salton proposed the so-called vector space representation for statistical modeling of text data, for a review see [16]. A term set is chosen and a document is represented by the vector of term frequencies. A document database then forms a so-called term-document matrix. The vector space representation can be used for classification and retrieval by noting that similar documents are somehow expected to be 'close' in the vector space. A simple Euclidean distance metric can be used if document vectors are properly normalized, otherwise angular distance may be used. This approach is principled, fast, and language independent. Deerwester and co-workers developed the concept of latent semantics based on PCA of the term-document matrix [15]. The fundamental observation behind the LSI approach is that similar documents use similar vocabularies, hence, the term vectors of a given topic could appear as produced by a stochastic process with highly correlated term-entries. By projecting the term-frequency vectors on a relatively low dimensional subspace, determined by the maximal amount of variance one would be able to filter out the inevitable 'noise'. Noise should here be thought of as individual document differences in term usage within a specific context. For well-defined topics, one could simply hope that a given context would have a stable core term set that would come out as a 'direction' in the term vector space. Below we will explain why this is likely not to happen in general document databases, and LSI is therefore often used as a dimensionality reduction tool, which is then post-processed to reveal cognitive components, e.g., by interactive visualization schemes [17].

2.2. Independent component analysis

Blind signal separation is the general problem of recovering source signals from an unknown mixture. This aim is in general not feasible without additional information. If we assume that the unknown mixture is linear and the sources are statistically independent processes, it is often possible to recover sources and mixing, using a variety of ICA techniques [1]. Here we will discuss some basic characteristics of mixtures and the possible recovery of sources.

First, we note that LSI/PCA is not able to reconstruct the mixing. PCA, being based on covariance is simply not informed enough to solve the problem. To see this let the mixture be given as

$$\mathbf{X} = \mathbf{AS}, \qquad X_{j,t} = \sum_{k=1}^{K} A_{j,k} S_{k,t}, \tag{1}$$

where $X_{j,t}$ is the value of j'th feature in the t'th measurement, $A_{j,k}$ is the mixture coefficient linking feature j with the component k, while $S_{k,t}$ is the level of activity in the k'th source. In a text instance a feature is a term and the measurements are documents, while the components can be interpreted as topical contexts.

As a linear mixture is invariant to an invertible linear transformation we need to define a normalization of one of the matrices **A**, **S**. We do this by assuming that the sources are unit variance. As they are assumed independent the covariance will thus be trivially given as the unit matrix. LSI, hence PCA, of the measurement matrix is based on analysis of the covariance

$$\Sigma_X = \lim_{T\to\infty} \frac{1}{T} \mathbf{X}\mathbf{X}^{\mathrm{T}} = \mathbf{A}\mathbf{A}^{\mathrm{T}} \tag{2}$$

Clearly the information in $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ is not enough to uniquely identify **A**, since if one solution **A** is found, any (row) rotated matrix $\tilde{\mathbf{A}} = \mathbf{AU}$, $\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}$ is also a solution, because $\tilde{\mathbf{A}}$ has the same outer product as **A**.

This is a potential problem for LSI based analysis. If the document database can be modeled as in (1) then the original characteristic context histograms will not be found by LSI. The field of ICA has on the other hand devised many algorithms that use more informed statistics to locate **A** and thus **S**, see [1] for a recent review.

The histogram of a source signal can roughly be described as sparse, normal, or dense. Scatter plots of projections of mixtures drawn from source distributions with one of these three characteristics are shown in Fig. 1. In the upper panel of Fig. 1, we show the typical appearance of a sparse source mixture. The sparse signal consists of relatively few large magnitude samples in a background of a large number of small signals. When mixing such independent sparse signals as in (1), we obtain a set of 'rays'

emanating from the origin. The directions of the rays are given by the column vectors of the **A**-matrix. If the sources are normally distributed (middle panel of Fig. 1) there is no additional information but the covariance matrix. Hence, in some sense this is a worst case for separation. Fortunately, many interesting real world data sets are very sparse, hence, more similar to the upper panel of Fig. 1.

3. Component analysis of speech

In the authoritative textbook 'Discrete-Time Processing of Speech Signals' by Deller et al. [18] the phoneme is defined as the class of sounds that are consistently perceived as representing a certain minimal linguistic unit. In American English approximately 40 phonemes are in use, of which 12 are vowels. Vowels vary in temporal duration between 40-400msec [18].

The processes in the speech production system are generally considered stationary for time intervals on the order of 20 msec [18], hence, we will use an analysis window of this duration. In each window we represent the sound signal, i.e., 200 signal values for a sampling rate of 10 kHz, by a relatively low-dimensional feature vector. This feature vector is obtained by homomorphic filtering, as often invoked in speech recognition. The resulting, so-called *cepstral coefficients* are designed to reduce the influence of the speech pitch, i.e., the speaker's 'tone' [18]. The cepstral coefficients are used in speaker independent speech recognition, because in this context the pitch is a confound. The speaker dependent and speaker independent aspect are separated in the cepstral coefficient representation, hence, we use this representation to emphasize the linguistic content and suppress the speakers 'voice print'.

A small set of four simple utterances ('s', 'o', 'f', 'a') from the TIMIT database [19] were used for this demonstration. For the analysis we used 20 msec analysis windows with 50% overlap. The windows were represented by 16 cepstral coefficients. The temporal development of the cepstral representation of the four utterances is presented in two versions in Fig. 2, in the upper panel for the training set, and in the lower panel for a test set. After variance normalization we sparsified the coefficients by zeroing windows of normalized magnitudes with a statistical $z < 1.7$. In Fig. 3 we show the scatter plot of the set of windows projected onto the first two principal components derived from the 16 x 16 sparsified feature covariance matrix. There is a marked 'ray' structure with rays emanating from the origin of the coordinate system (0,0). The projected features from the set of analysis windows have been annotated with their utterance

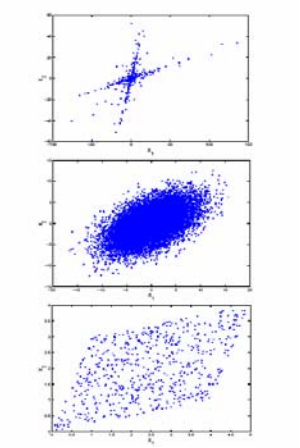

Fig. 1. Prototypical feature distributions
Prototypical feature distributions produced by a linear mixture, based on sparse (top), normal (middle), or dense source signals (bottom), respectively. The characteristic of the sparse signal is that it consists of relatively few large magnitude samples on a background of small signals.

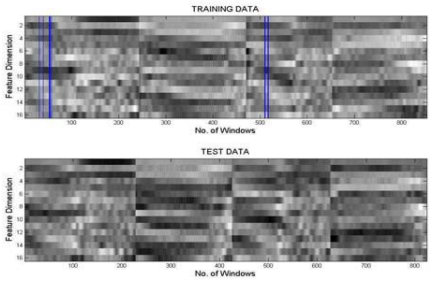

Fig. 2. Cepstral coefficient sequences for training and test sets
Four separate utterances are concatenated for this experiment, representing the sounds 's', 'o', 'f', 'a'. Each concatenated set of utterances is represented twice: in a training set and in a test set. The boundaries between the four utterances are clearly visible, and we note that the utterances show much similarity between the two samples (test and train), however, they are of quite different duration. The first of the two phones of the utterance 's' is the opening a-like phoneme. In the upper panel we have added a set of vertical lines to indicate positions of analysis windows that belong to a generalizable finger print feature further discussed in Fig. 3.

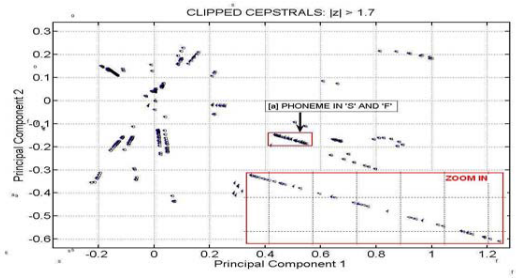

Fig. 3. Scatter plot of data on latent space

We show the latent space formed by the two first principal components of the training data consisting of four separate utterances shown in figure 2 representing the sounds 's', 'o', 'f', 'a'. The structure clearly resembles the sparse component mixture in Fig. 1, with 'rays' emanating from the origin (0,0). The ray marked with an arrow contains a mixture of 's' and 'f' analysis windows. The locations of these windows were indicated by vertical lines in Fig. 2. This feature also contains a mixture of windows from both the training and test utterances, hence, is a generalizable characteristic feature associated with the vowel a-like sound that opens both an 's' and an 'f'.

origin. The arrow points to a linear ray structure which contains windows from utterances 's' and 'f'. In order to understand which part of the utterances these windows belong to, we have marked up several points (windows) in Fig. 3 and have indicated the temporal location of these windows as vertical stripes in Fig. 2. It is clear that the feature is related to the similar a-like sound that opens both 's' and 'f'. The generalizability of this structure was proved by creating a similar plot with the projections of the *test set* windows (data not shown). This structure is indeed generalizable in contrast to some of the other ray-like structures that apparently are too specific to provide generalization from the relative small set of training data.

The results seem to indicate that generalizable cognitive components corresponding to phonemes can be identified using linear component analysis. The ray structures representing the phonemes are not aligned with the directions of the principal components, hence, an ICA scheme is required. Phoneme recognition is an active research field in speech recognition, see e.g., [20], and it is an interesting issue for further research whether the generalizable structure found in this work can assist phoneme recognition in general.

4. Voice print components

While phonemes are universal components of language and generalizable in large populations, *speaker identity* plays an important role both in social contexts and in speech based engineering applications, e.g., related to access control [21].

Speaker recognition has two aspects: Speaker identification, and speaker verification. Speaker verification is the process of determining whether a postulated speaker identity is correct, while speaker identification is the process of finding the identity of an unknown speaker by comparing his/her voice with all the registered/known speakers in the database [22]. In the case that the unknown speaker must come from a fixed set of enrolled speakers, the system is referred to as a closed-set system. Speaker recognition systems are moreover divided according to the spoken text modality: text-dependent and text-independent. Compared to text-dependent speaker recognition, text-independent systems are more flexible, but also more complex. The most widely accepted features for speaker recognition are mel-frequency cepstral coefficients (MFCC). The MFCCs are perceptually weighted cepstral coefficients [18].

According to our basic hypothesis the speaker dependent generalizable 'cognitive' components should be elucidated by Latent Semantic Indexing (LSI). To test the hypothesis we study here three speakers' voice messages from our in-house ELSDSR speech database [23]. In this database, read text is recorded using a MARANTZ PMD670 portable solid state recorder, and stored in PCM (wav) format. The sampling frequency is 16 kHz. ELSDSR contains voice messages from a total of 22 speakers (12M/10F) of age from 24y to 63y.

Speaker identity information in speech can be categorized into a hierarchy ranging from low-level cues, such as the basic sound of a person's voice, which is related to physical traits of the vocal apparatus, to high-level cues, such as particular word usage (idiolect), conversational patterns and even topics of conversations, which is related to learned habits and style [24].

For the first *text-dependent* speaker recognition experiment, signals from speakers F1, F2 and M1 reading the same text content were selected, and divided into training set (52.5sec) and test set (35.5sec). The windows with 20 msec signal content were blocked without overlap, and 12 MFCCs were extracted from each window. To form the long-term features, 50 basic analysis windows were concatenated. The dimensionality of the aggregate representation is thus 50 x 12. The total number of such expanded windows in the analysis was 522. After variance normalization, energy based sparsification was performed on the high dimensional data, and the upper 1% fraction was retained. Finally, LSI (PCA) was performed on the sparsified data to get the scatter plot of the data on the subspace spanned by three latent dimensions (LD), shown in Fig. 4. We annotated the data points for the training set of the three speakers as: F1 (red square), F2 (blue diamond) and M1 (black x); and test set as: F1 (cyan +), F2 (green triangle) and M1 (magenta circle). Since the speakers read the same text content (training and test set are different) the red, blue and black points emanate from (0,0), and show similar sparse ICA 'ray' structures. These features of same text also carry characteristics of the given words, i.e., similar to the phoneme features found above. However, importantly the rays also show speaker-dependent characteristics. This is most easily appreciated by inspecting the three plots to the right in Fig. 4. Here the situations for the individual speaker are depicted as seen, the features do not generalize in a simple way, it appears that there is an offset between test data and training data, which is speaker dependent. We therefore stipulate that this effect is an interaction between the text content and the speaker identity.

We now turn to text-independent speech. We study the same three speakers as before, two female and one male. The representation is identical to the one used for the text-dependent experiment. The scatter plot of test and training data is shown in 3D subspace based on latent dimensions 2nd, 4th and 5th. Fig. 5 shows that data points from 2 female speakers and the male speaker are aligned for both training and test set. The right side panel shows a zoomed in and projected subset of the data belonging to the two female speakers in latent dimension 4 and 5. Thus the generalizable ray structure emanates from (0,0) *without* offsets.

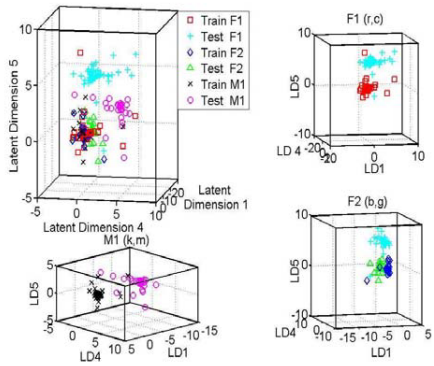

Fig. 4. Text-dependent speaker recognition
We focus on text-dependent speech. The basic analysis window of the speech signal is represented by 12 MFCCs. 50 basic analysis windows are concatenated to form an intermediate time scale representation. We sparsified the coefficients by retaining the upper 1% magnitude fraction. We used a training set from speakers F1, F2 and M1. The data from the training set is submitted for LSI, we show the scatter plots of both training and test data in the space of the 1st, 4th and 5th latent components. The upper left display shows all data points. There is an evident ray structure corresponding to a generative ICA model based on linear mixing of sparse sources, i.e., similar to the situation seen at the basic time scale analysis window (20 msec). The structure is indeed speaker dependent in the sense that the ray systems are offset from the origin. We conclude that we find a mixture of phoneme like features and speaker identity features.

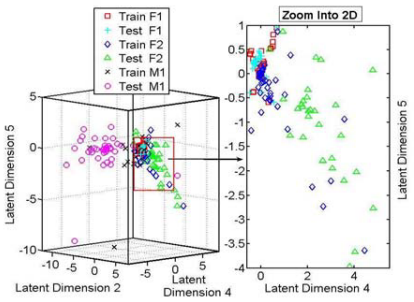

Fig. 5. Text-independent speaker recognition
We focus on text-independent speech. The setup is the same as text-dependent case. In the left panel all data points are shown as represented in the space of the 2nd, 4th and 5th latent components. There is an evident ray structure corresponding to a generative ICA model based on linear mixing of sparse sources. In contrast to the text-dependent case we see that the ray structure is solely determined by the speaker identity. The right hand side plot shows a close up of the structure for the female speaker F2: emphasizing the generalizability. The rays from the training and test sets are closely aligned.

5. Conclusion

We have proposed to define cognitive component analysis as the process of unsupervised grouping of data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity. In this paper we have studied the derived cognitive components of speech signals. We used homomorphic filtering to derive features, and analyzed the excursion set after thresholding based on energy.

At short time scales, we found generalizable features corresponding to phonemes. Phonemes are universal linguistic atoms recognized by large populations. Humans swiftly and reliably recognize other human's voice. We have shown that at intermediate time scales, 500-1000msec, there are generalizable speaker specific sparse components.

The fact that we find such cognitively relevant component by simple unsupervised learning based on sparse linear component analysis lends further support to our working hypothesis that humans could use such information theoretical representations, not only in basic perception tasks, but also when analyzing more abstract data.

6. Acknowledgment

This work is supported by the Danish Technical Research Council, through the framework project 'Intelligent Sound', www.intelligentsound.org (STVF No. 26-04-0092).

References

[1] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[2] Anthony J. Bell and Terrence J. Sejnowski, "The 'independent components' of natural scenes are edge filters," *Vision Research*, vol. 37, no. 23, pp. 3327–3338, 1997.

[3] Patrik Hoyer and Aapo Hyvrinen, "Independent component analysis applied to feature extraction from colour and stereo images," *Network: Comput. Neural Syst.*, vol. 11, no. 3, pp. 191–210, 2000.

[4] M.S. Lewicki, "Efficient coding of natural sounds," *Nature Neuroscience*, vol. 5, no. 4, pp. 356–363, 2002.

[5] L. K. Hansen, P. Ahrendt, and J. Larsen, "Towards cognitive component analysis," in *AKRR'05 -International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*. Jun 2005, Pattern Recognition Society of Finland, Finnish Artificial Intelligence Society, Finnish Cognitive Linguistics Society.

[6] C.M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, 1995.

[7] L. K. Hansen, J. Larsen, and T. Kolenda, "On independent component analysis for multimedia signals," in *Multimedia Image and Video Processing*, pp. 175–199. CRC Press, Sep 2000.

[8] L. K. Hansen, J. Larsen, and T. Kolenda, "Blind detection of independent dynamic components," in *IEEE International Conference on Acoustics, Speech, and Signal Processing 2001*, 2001, vol. 5, pp. 3197–3200.

[9] T. Kolenda, L. K. Hansen, and J. Larsen, "Signal detection using ICA: Application to chat room topic spotting," in *Third International Conference on Independent Component Analysis and Blind Source Separation*, 2001, pp. 540–545.

[10] T. Kolenda, L. K. Hansen, J. Larsen, and O.Winther, "Independent component analysis for understanding multimedia content," in *Proceedings of IEEE Workshop on Neural Networks for Signal Processing XII*, H. Bourlard et al. Ed., Piscataway, New Jersey, 2002, pp. 757–766, IEEE Press, Martigny, Valais, Switzerland, Sept. 4-6, 2002.

[11] J. Larsen, L.K. Hansen, T. Kolenda, and F.AA. Nielsen, "Independent component analysis in multimedia modeling," in *Fourth International Symposion on Independent Component Analysis and Blind Source Separation*, Shun ichi Amari et al. Ed., Nara, Japan, apr 2003, pp. 687–696, Invited Paper.

[12] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[13] Pedro A. D. F. R. Højen-Sørensen, Ole Winther, and Lars Kai Hansen, "Mean-field approaches to independent component analysis," *Neural Comput.*, vol. 14, no. 4, pp. 889–918, 2002.

[14] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, 1973.

[15] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman, "Indexing by latent semantic analysis," *JASIS*, vol. 41, no. 6, pp. 391–407, 1990.

[16] Gerard Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, 1989.

[17] T.K. Landauer, D. Laham, and M. Derr, "From paragraph to graph: latent semantic analysis for information visualization," *Proc Natl Acad Sci*, vol. 101, no. Sup. 1, pp. 5214–5219, 2004.

[18] John R. Deller, John H. Hansen, and John G. Proakis, *Discrete Time Processing of Speech Signals*, IEEE Press Marketing, 2000.

[19] J. S. Garofolo et al., *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM*, NIST, 1993.

[20] Ofer Dekel, Joseph Keshet, and Yoram Singer, "An online algorithm for hierarchical phoneme classification," in *MLMI*, 2004, pp. 146–158.

[21] *http://www.research.ibm.com/VIVA Demo*, 2005.

[22] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *ICASSP 2002*, 2002.

[23] *http://www.imm.dtu.dk/~lf/ELSDSR.htm*, 2005.

[24] J.P. Campbell, D.A. Reynolds, and R.B. Dunn, "Fusing high- and low-level features for speaker recognition," in *Proceedings of Eurospeech-2003 (Geneva, Switzerland)*, 2003, pp. 2665–2668.

Appendix C

Phonemes as Short Time Cognitive Components

This article is published in *Proc. International Conference on Acoustics, Speech, and Signal Processing* 2006, vol. 5, pp 869-872, with the same title. Authors are Ling Feng and Lars Kai Hansen. It is also available as IMM publication database with number imm4058.

PHONEMES AS SHORT TIME COGNITIVE COMPONENTS

Ling Feng and Lars Kai Hansen
Informatics and Mathematical Modeling,
Technical University of Denmark, Denmark
lf@imm.dtu.dk, lkh@imm.dtu.dk

ABSTRACT

Cognitive component analysis (COCA) is defined as the process of unsupervised grouping of data such that the resulting group structure is well-aligned with that resulting from human cognitive activity [1]. In this paper we address COCA in the context short time sound features, finding phonemes which are the smallest contrastive units in the sound system of a language. Generalizable components were found deriving from phonemes based on homomorphic filtering features with basic time scale (20 msec). We sparsified the features based on energy as a preprocessing means to eliminate the intrinsic noise. Independent component analysis was compared with latent semantic indexing, and was demonstrated to be a more appropriate model in COCA.

1. INTRODUCTION

Cognitive component analysis (COCA) as a newly defined concept was first brought to bear in [1]: the process of unsupervised grouping of data such that the resulting group structure is well-aligned with that resulting from human cognitive activity. The concept is related to Lee and Seung's work on non-negative matrix factorization (NMF). In [2] they showed that components could be understood using concepts from gestalt theory: the factorization of an observation matrix in terms of a relatively small set of cognitive components leads to a parts-based object representation. In 2002, similar parts-based decompositions were obtained in a latent variable model based on non-negative linear mixtures of non-negative independent source signals [3]. Holistic, but parts-based, recognition of objects is frequently reported in perception studies across multiple modalities and increasingly in abstract data, where object recognition is a cognitive process.

The human perceptual system can model complex multi-agent scenery by using a broad spectrum of cues for analyzing perceptual input and for identification of individual signal producing agents. The fact motivating our interest in COCA is that representations found in human and animal perceptual systems closely resemble the theoretically optimal representations from the unsupervised signal separation, namely independent component analysis (ICA) [4, 5, 6]. This paper further discusses the generality of COCA based on the previous work [1, 7], and tries to answer the question: *Are such optimal representations based on abstract "independence" also relevant in higher cognitive functions?*

The phoneme is the smallest contrastive unit in the sound system of a language. Phoneme recognition is an active research field in speech recognition, see e.g., [8]. In [7] phonemes have been investigated by one of the generic tools of COCA analysis, namely Latent Semantic Indexing (LSI), and generalizable components and structures representing some of these smallest units have been found, as illustrated in Fig. 1. However whether the generalizable structure found in this work can assist phoneme recognition in general, still needs to be explored. Grouping by ICA has been pursued earlier for several abstract data types including text, dynamic text (chat), images, and combinations [9, 10, 11, 12, 13]. It was found that ICA is a more appropriate model than both LSI, which is too constrained, and clustering, which may in some instances be too flexible as a representation of text data.

The generality of ICA makes it possible to be utilized in many different areas. The classical application in signal processing of ICA model is blind source separation (BSS). A classical example of BSS is the cocktail party problem (CPP), see e.g., [14]. The problem is to separate the voices of different speakers, using recordings of one or more microphones. Comparing to BSS/CPP which is basically using original sound signals, the ICA model in COCA analysis applies on homomorphic filtering features, namely Mel-frequency Cepstral Coefficient (MFCC). MFCCs are short-term spectral features, and the mel-frequency warping transformation based on human auditive system. In COCA we are interested in a cognitive level, so to speak before semantics. The features we look for can be compared to the features a foreign speaker hears on entry. Sounds are recognized but without semantic reference. Hence, the cognitive context in our COCA is in the intermediate-level between source separation (low-level) and content recognition (high-level).

2. COGNITIVE COMPONENT ANALYSIS

2.1 Latent semantic indexing (LSI)

Latent semantic indexing is the PCA applied on abstract data such as text [15]. It is basically a tool for dimensionality reduction and also can be used to find group structure in data when the signal-to-noise ratio is high [7]. Our approach is inspired by LSI and the main innovation here is the active search for generalizable non-orthogonal linear features that may be described in terms of an independent component generative model.

A strong assumption in LSI is that the data have Gaussian distribution. Unfortunately, many real world data are *nongaussian*, instead very sparse [1, 7]. Hence LSI is often used as a tool to reduce dimensionality, which is post-processed to reveal cognitive components, e.g., by interactive visualization schemes [16].

2.2 Independent component analysis (ICA)

ICA algorithms can estimate independent components from linear mixtures [17], and has applications in many real world data. Here we discuss some basic characteristics of mixtures and the possible recovery of sources.

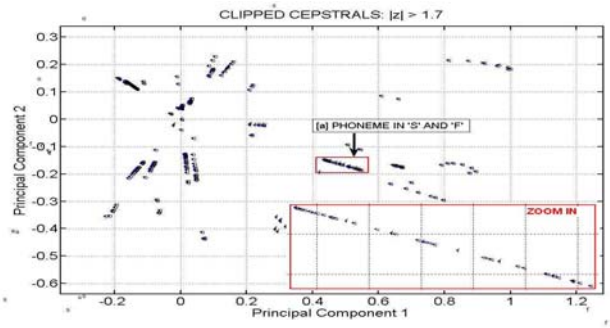

Fig. 1. Scatter plot of data on latent space

The latent space is formed by the two first principal components of the training data consisting of four separate utterances representing the sounds 's', 'o', 'f', 'a'. The structure clearly shows the sparse component mixture, with 'rays' emanating from the origin (0,0). The ray marked with an arrow contains a mixture of 's' and 'f' analysis windows, a generalizable characteristic feature associated with the vowel a-like sound that opens both an 's' and an 'f'.

First, we note that LSI/PCA is not able to reconstruct the mixing. PCA, being based on co-variance is simply not informed enough to solve the problem. To see this let the mixture be given as

$$\mathbf{X} = \mathbf{AS}, \qquad X_{j,t} = \sum_{k=1}^{K} A_{j,k} S_{k,t}, \tag{1}$$

where $X_{j,t}$ is the value of j'th feature in the t'th measurement, $A_{j,k}$ is the mixture coefficient linking feature j with the component k, while $S_{k,t}$ is the level of activity in the k'th source. In a text instance a feature is a term and the measurements are documents, while the components can be interpreted as topical contexts.

As a linear mixture is invariant to an invertible linear transformation we need to define a normalization of one of the matrices **A**, **S**. We do this by assuming that the sources are unit variance. As they are assumed independent the covariance will thus be trivially given as the unit matrix. LSI, hence PCA, of the measurement matrix is based on analysis of the covariance

$$\Sigma_X = \lim_{T\to\infty} \frac{1}{T} \mathbf{X}\mathbf{X}^{\mathrm{T}} = \mathbf{A}\mathbf{A}^{\mathrm{T}} \tag{2}$$

Clearly the information in $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ is not enough to uniquely identify **A**, since if one solution **A** is found, any (row) rotated matrix $\widetilde{\mathbf{A}} = \mathbf{AU}$, $\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}$ is also a solution, because $\widetilde{\mathbf{A}}$ has the same outer product as **A**. This is a potential problem for LSI based analysis. The ICA community has on the other hand devised many algorithms that use more informed statistics to locate **A** and thus **S**, see [17] for a recent review.

3. COMPONENT ANALYSIS FOR PHONEMES

The phoneme is defined as the class of sounds that are consistently perceived as representing a certain minimal linguistic unit in [18]. However phonologists have differing views of the phoneme, and two major ones are: in the American structuralist tradition, a phoneme is defined according to its allophones and environments; in the generative tradition, a phoneme is defined as a set of distinctive features [19]. An allophone is a phonetic variant of a phoneme in a particular language. According to the first view, the same phoneme can sound slightly different in different languages and environments. In American English approximately 40 phonemes are in use, of which 12 are vowels. Vowels vary in temporal duration between 40-400msec [18].

Four simple utterances 's', 'o', 'f', 'a' from the TIMIT database [20] were used for this demonstration. The basic time scale of 40 msec was used (windowing with 95% overlap), since the speech production system is generally considered stationary for time intervals on the order of 20-40 msec [18]. The windows were represented by 16 MFCCs. The temporal development of the mel-cepstral representation of the four utterances is presented in the upper panel of Fig. 4. After variance normalization we sparsified the energy based coefficients by zeroing windows of normalized magnitudes with a statistical $z < 1.4$, which retains 55% energy from original features. LSI/PCA was performed on the sparsified feature coefficients to get the most variant PCA components. The results from Fig. 1 seem to indicate that generalizable cognitive components corresponding to phonemes, e.g. /æ/ from utterance 's' and 'f', can be identified using linear component analysis. However the ray structures representing the phonemes are not aligned with the directions of the principal components, hence, an ICA scheme is required.

Six components ICA was applied on the PCA coefficients. Fig. 2 shows the scatter plot of sparsified features on the first two principal components derived from the 16 x 16 sparsified feature covariance matrix. The six independent sources were annotated as red circle, blue square, green diamond, magenta +, cyan triangle and black X respectively. The tags for the samples were labeled according to the independent sources, **S** matrix**,** from ICA analysis on sparsified and dimensionality reduced features. The arrows in Fig. 2 represent the directions of sources which are the column vectors of the mixing matrix **A** in equation (1). The 'ray' structure with rays emanating from the origin of the coordinate system is evident, and each ray along the vector belongs to one independent source. In order to testify the generalizability of this structure, a test set with another set of utterances 's', 'o', 'f', 'a' from TIMIT

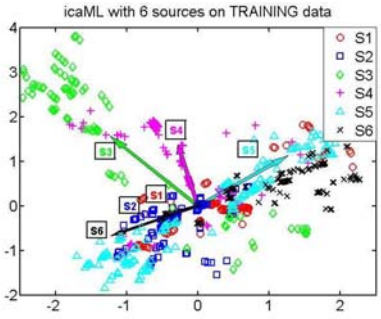

Fig. 2. Scatter plot of training data
Six components ICA performed on PCA coefficients. Scatter plot shows the data projected on the first two principal components derived from the sparsified features. The circle, square, diamond, +, triangle and X stand for 6 independent sources. The tags for the samples were labeled according to **S** matrix from ICA, and the arrows represent the directions of sources from mixing matrix **A**. The 'ray' structure with rays emanating from the origin (0,0) is evident.

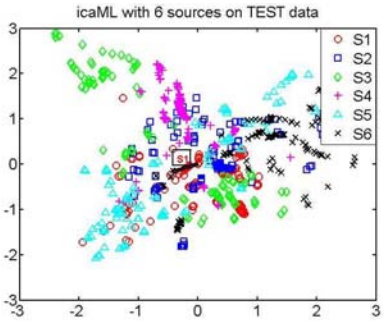

Fig. 3. Scatter plot of test data
Another set of utterances 's', 'o', 'f', 'a' was analyzed. The 'ray' structure is obvious and similar to the training set, emanating from the origin (0,0).

was analyzed using the same setup. The results are shown in Fig. 3. Here we only show the direction of the first source. Later we will demonstrate the cognitive content of this source.

Generalizability has been verified in another way by using two different implementations of ICA, namely maximum likelihood ICA (icaML) and the fast fixed-point algorithm for ICA (fastICA). IcaML algorithm is the estimation of the independent component as in the Infomax by Bell and Sejnowski [21] using a maximum likelihood formulation. Fig. 4 and 5 show the classification results from icaML and fastICA on training and test sets separately. In the two upper panels, the temporal development

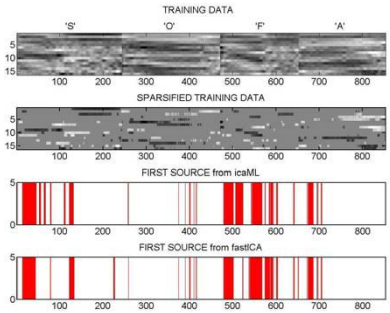

Fig. 4. MFCCs and Classification on Training set
In the two upper panels, the temporal development of the mel-frequency cepstral representations of the original 's', 'o', 'f', 'a' and 4 spasified ones is presented. The boundaries between them are clearly visible. 55% energy was retained after sparsification. The first independent sources from two ICA implementations are shown in the two lower panels: the vertical lines indicate the locations of windows belonging to the first source. Results from two ICA algorithms are similar. A large percentage of the windows locate in, approximately, windows No. 1 to No. 133 for 's', and No. 471 to No. 600 for 'f'. It indicates the feature is related to the similar /æ/ sound that opens both 's' and 'f'.

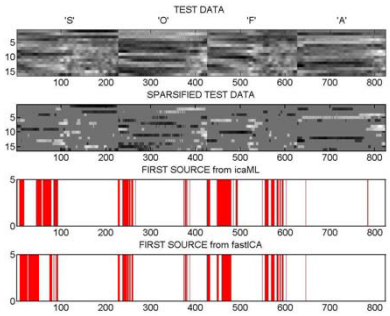

Fig. 5. MFCCs and Classification on Test set
The two upper panels show the temporal development of the mel-frequency cepstral representations of the four original utterances and four spasified ones. 60% energy was left for test set. The two lower panels show the first independent sources from icaML and fastICA: the vertical lines indicate the locations of windows belonging to the first source. Two panels look quite similar. The similar scenario shown in Fig. 4 for training set happened again on test set, which indicates the feature is related to the similar /æ/ sound that opens both 's' and 'f'. However there are more mis-detections located outside the above ranges.

of the mel-frequency cepstral representations of the four original utterances and four sparsified utterances is presented with the sequence of 's', 'o', 'f', 'a'. The boundaries between the four utterances are clearly visible, and the utterances show much similarity between the two samples (test and train), however, they are of quite different duration. For training set, 55% energy was retained after sparsification; and 60% energy was left for test set. The first independent sources from two ICA algorithms are shown in the two lower panels of Fig. 4 and 5: the vertical lines indicate the locations of windows belonging to the first source. It is quite clear that the results of icaML resemble those of fastICA. For training set, we notice that a large percentage of the windows locate in the first part of 's' and 'f' utterances, which approximately from windows No. 1 to No. 133 for 's', and No. 471 to No. 600 for 'f'. It indicates the feature is related to the similar /æ/ sound that opens both 's' and 'f'. A similar scenario happened in test set, however there are more lines locate outside the above ranges. Our interpretation is the windows containing low energy (almost zero) have simply been classified into the first class. The classification has been improved while we slightly reduced the threshold for sparsification. However low threshold brings more noise, which increases the classification error.

4. CONCLUSION

The generality of cognitive component analysis, which is defined as the process of unsupervised grouping of data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity, has been explored in this paper. We posit speech COCA in a cognitive level before semantics. In other words, sounds (sources) are recognizable, but without semantic reference. Therefore COCA is localized in the intermediate-level between source separation (low-level) and content recognition (high-level).

We have studied the derived cognitive components of phonemes from short time homomorphic filtering features with energy based sparsification. ICA on short-term spectral features, MFCC, was compared with latent semantic indexing, and was demonstrated to be a more appropriate model in COCA.

The fact that we find the 'ray' structure of cognitively relevant components by simple unsupervised learning based on sparse linear component analysis highlights the possibility of using unlabeled samples in supervised learning.

5. ACKNOWLEDGMENT

This work is supported by the Danish Technical Research Council, through the framework project 'Intelligent Sound', www.intelligentsound.org (STVF No. 26-04-0092).

REFERENCES

[1] L. K. Hansen, P. Ahrendt, and J. Larsen, "Towards cognitive component analysis," in *AKRR'05 -International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*. Jun 2005, Pattern Recognition Society of Finland, Finnish Artificial Intelligence Society, Finnish Cognitive Linguistics Society.

[2] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[3] Pedro A. D. F. R. Højen-Sørensen, Ole Winther, and Lars Kai Hansen, "Mean-field approaches to independent component analysis," *Neural Comput.*, vol. 14, no. 4, pp. 889–918, 2002.

[4] Anthony J. Bell and Terrence J. Sejnowski, "The 'independent components' of natural scenes are edge filters," *Vision Research*, vol. 37, no. 23, pp. 3327–3338, 1997.

[5] Patrik Hoyer and Aapo Hyvrinen, "Independent component analysis applied to feature extraction from colour and stereo images," *Network: Comput. Neural Syst.*, vol. 11, no. 3, pp. 191–210, 2000.

[6] M.S. Lewicki, "Efficient coding of natural sounds," *Nature Neuroscience*, vol. 5, no. 4, pp. 356–363, 2002.

[7] L. Feng and L. K. Hansen, "On low level cognitive components of speech," *accepted* in CIMCA'05 -International Conference on Computational Intelligence for Modelling, Nov 2005.

[8] Ofer Dekel, Joseph Keshet, and Yoram Singer, "An online algorithm for hierarchical phoneme classification," in *MLMI*, pp. 146–158, 2004.

[9] L. K. Hansen, J. Larsen, and T. Kolenda, "On independent component analysis for multimedia signals," in *Multimedia Image and Video Processing*, pp. 175–199. CRC Press, Sep 2000.

[10] L. K. Hansen, J. Larsen, and T. Kolenda, "Blind detection of independent dynamic components," in *IEEE International Conference on Acoustics, Speech, and Signal Processing 2001*, vol. 5, pp. 3197–3200, 2001.

[11] T. Kolenda, L. K. Hansen, and J. Larsen, "Signal detection using ICA: Application to chat room topic spotting," in *Third International Conference on Independent Component Analysis and Blind Source Separation*, pp. 540–545, 2001.

[12] T. Kolenda, L. K. Hansen, J. Larsen, and O.Winther, "Independent component analysis for understanding multimedia content," in *Proceedings of IEEE Workshop on Neural Networks for Signal Processing XII*, H. Bourlard et al. Ed., Piscataway, New Jersey, 2002, pp. 757–766, IEEE Press, Martigny, Valais, Switzerland, Sept. 4-6, 2002.

[13] J. Larsen, L.K. Hansen, T. Kolenda, and F.AA. Nielsen, "Independent component analysis in multimedia modeling," in *Fourth International Symposion on Independent Component Analysis and Blind Source Separation*, Shun ichi Amari et al. Ed., Nara, Japan, apr 2003, pp. 687–696, Invited Paper.

[14] S. Haykin and Z. Chen, "The Cocktail Party Problem," Neural Comp. 2005, vol. 17, pp. 1875-1902.

[15] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman, "Indexing by latent semantic analysis," *JASIS*, vol. 41, no. 6, pp. 391–407, 1990.

[16] T.K. Landauer, D. Laham, and M. Derr, "From paragraph to graph: latent semantic analysis for information visualization," *Proc Natl Acad Sci*, vol. 101, no. Sup. 1, pp. 5214–5219, 2004.

[17] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[18] John R. Deller, John H. Hansen, and John G. Proakis, *Discrete Time Processing of Speech Signals*, IEEE Press Marketing, 2000.

[19] E. E. Loos, S. Anderson, D. H. Jr. Day, P. C. Jordan and J. D. Wingate, "Glossary of linguistic terms," SIL International, 2004. http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/index.htm

[20] J. S. Garofolo et al., *DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM*, NIST, 1993.

[21] A. Bell and T.J. Sejnowski, "An Information-Maximization Approach to Blind Separation and Blind Deconvolution," Neural Comp. 1995, vol.7, pp. 1129-1159.

Appendix D

Cogito Componentiter Ergo Sum

This article is accepted for publication in *Proc. International Conference on Independent Component Analysis and Blind Source Separation* 2006, pp 446-453, with the same title. Authors are Lars Kai Hansen and Ling Feng. It is also available as IMM publication database with number imm4141.

Cogito componentiter ergo sum

Lars Kai Hansen and Ling Feng

Informatics and Mathematical Modelling,
Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark
lkh,lf@imm.dtu.dk, www.imm.dtu.dk

Abstract. Cognitive component analysis (COCA) is defined as the process of unsupervised grouping of data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity. We present evidence that independent component analysis of abstract data such as text, social interactions, music, and speech leads to low level cognitive components.

1 Introduction

During evolution human and animal visual, auditory, and other primary sensory systems have adapted to a broad ecological ensemble of natural stimuli. This long-time on-going adaption process has resulted in representations in human and animal perceptual systems which closely resemble the information theoretically optimal representations obtained by independent component analysis (ICA), see e.g., [1] on visual contrast representation, [2] on visual features involved in color and stereo processing, and [3] on representations of sound features. For a general discussion consult also the textbook [4]. The human perceptional system can model complex multi-agent scenery. Human cognition uses a broad spectrum of cues for analyzing perceptual input and separate individual signal producing agents, such as speakers, gestures, affections etc. Humans seem to be able to readily adapt strategies from one perceptual domain to another and furthermore to apply these information processing strategies, such as, object grouping, to both more abstract and more complex environments, than have been present during evolution. Given our present, and rather detailed, understanding of the ICA-like representations in primary sensory systems, it seems natural to pose the question: *Are such information optimal representations rooted in independence also relevant for modeling higher cognitive functions?* We are currently pursuing a research programme, trying to understand the limitations of the ecological hypothesis for higher level cognitive processes, such as grouping abstract objects, navigating social networks, understanding multi-speaker environments, and understanding the representational differences between self and environment.

Wagensberg has pointed to the importance of independence for successful 'life forms' [5]

> A living individual is part of the world with some identity that tends to become independent of the uncertainty of the rest of the world

Thus natural selection favors innovations that increase independence of the agent in the face of environmental uncertainty, while maximizing the gain from the predictable aspects of the niche. This view represents a precision of the classical Darwinian formulation that natural selection simply favors adaptation to given conditions. Wagensberg points out that recent biological innovations, such as nervous systems and brains are means to decrease the sensitivity to un-predictable fluctuations. An important aspect of environmental analysis is to be able to recognize event induced by the self and other agents. Wagensberg also points out that by creating alliances agents can give up independence for the benefit of a group, which in turns may increase independence for the group as an entity. Both in its simple one-agent form and in the more tentative analysis of the group model, Wagensberg's theory emphasizes the crucial importance of *statistical independence* for evolution of perception, semantics and indeed cognition. While cognition may be hard to quantify, its direct consequence, human behavior, has a rich phenomenology which is becoming increasingly accessible to modeling. The digitalization of everyday life as reflected, say, in telecommunication, commerce, and media usage allows quantification and modeling of human patterns of activity, often at the level of individuals. Grouping of events or objects in categories is

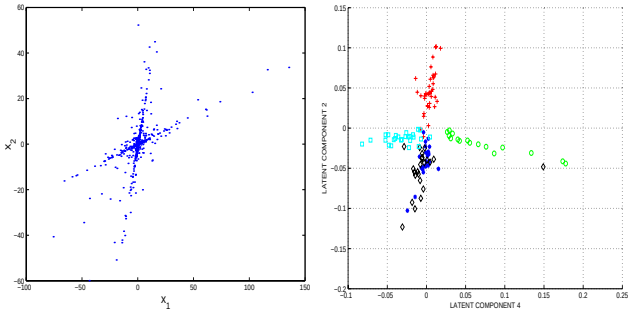

Fig. 1. Generic feature distribution produced by a linear mixture of sparse sources (left) and a typical 'latent semantic analysis' scatter plot of principal component projections of a text database (right). The characteristics of a sparse signal is that it consists of relatively few large magnitude samples on a background of small signals. Latent semantic analysis of the so-called MED text database reveals that the semantic components are indeed very sparse and does follow the laten directions (principal components). Topics are indicated by the different markers. In [6] an ICA analysis of this data set post-processed with simple heuristic classifier showed that manually defined topics were very well aligned with the independent components. Hence, constituting an example of cognitive component analysis: Unsupervised learning leads to a label structure corresponding to that of human cognitive activity.

fundamental to human cognition. In machine learning, classification is a rather

well-understood task when based on *labelled* examples [7]. In this case classification belongs to the class of *supervised* learning problems. Clustering is a closely related *unsupervised* learning problem, in which we use general statistical rules to group objects, without a priori providing a set of labelled examples. It is a fascinating finding in many real world data sets that the label structure discovered by unsupervised learning closely coincides with labels obtained by letting a human or a group of humans perform classification, labels derived from human cognition. *We thus define cognitive component analysis (COCA) as unsupervised grouping of data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity* [8]. This presentation is based on our earlier results using ICA for abstract data such as text, dynamic text (chat), web pages including text and images, see e.g., [9–13].

2 Where have we found cognitive components?

Text analysis. Symbol manipulation as in text is a hallmark of human cognition. Salton proposed the so-called vector space representation for statistical modeling of text data, for a review see [14]. A term set is chosen and a document is represented by the vector of term frequencies. A document database then forms a so-called term-document matrix. The vector space representation can be used for classification and retrieval by noting that similar documents are somehow expected to be 'close' in the vector space. A metric can be based on the simple Euclidean distance if document vectors are properly normalized, otherwise angular distance may be useful. This approach is principled, fast, and language independent. Deerwester and co-workers developed the concept of latent semantics based on principal component analysis of the term-document matrix [15]. The fundamental observation behind the latent semantic indexing (LSI) approach is that similar documents are using similar vocabularies, hence, the vectors of a given topic could appear as produced by a stochastic process with highly correlated term-entries. By projecting the term-frequency vectors on a relatively low dimensional subspace, say determined by the maximal amount of variance one would be able to filter out the inevitable 'noise'. Noise should here be thought of as individual document differences in term usage within a specific context. For well-defined topics, one could simply hope that a given context would have a stable core term set that would come out as a eigen 'direction' in the term vector space. The orthogonality constraint of co-variance matrix eigenvectors, however, often limits the interpretability of the LSI representation, and LSI is therefore more often used as a dimensional reduction tool. The representation can be post-processed to reveal cognitive components, e.g., by interactive visualization schemes [16]. In Figure 1 (right) we indicate the scatter plot of a small text database. The database consists of documents with overlapping vocabulary but five different (high level cognitive) labels. The 'ray'-structure signaling a sparse linear mixture is evident.

Social networks. The ability to understand social networks is critical to humans. Is it possible that the simple unsupervised scheme for identification of independent components could play a role in this human capacity? To investigate this issue we have initiated an analysis of a well-known social network of some practical importance. The so-called *actor network* is a quantitative rep-

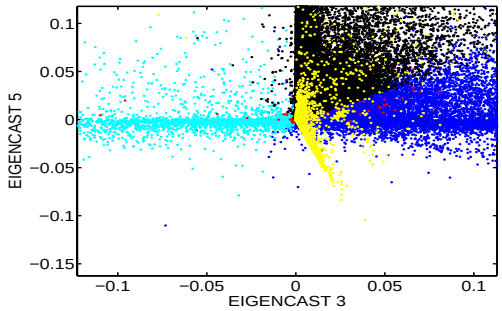

Fig. 2. The so-called actor network quantifies the collaborative pattern of 382.000 actors participating in almost 128.000 movies. For visualization we have projected the data onto principal components (LSI) of the actor-actor co-variance matrix. The eigenvectors of this matrix are called 'eigencasts' and they represent characteristic communities of actors that tend to co-appear in movies. The network is extremely sparse, so the most prominent variance components are related to near-disjunct sub-communities of actors with many common movies. However, a close up of the coupling between two latent semantic components (the region $\sim (0,0)$) reveals the ubiquitous signature of a sparse linear mixture: A pronounced 'ray' structure emanating from (0,0). The ICA components are color coded. We speculate that the cognitive machinery developed for handling of independent events can also be used to locate independent sub-communities, hence, navigate complex social networks.

resentation of the co-participation of actors in movies, for a discussion of this network, see e.g., [17]. The observation model for the network is not too different from that of text. Each movie is represented by the *cast*, i.e., the list of actors. We have converted the table of the about $T = 128.000$ movies with a total of $J = 382.000$ individual actors, to a sparse $J \times T$ matrix. For visualization we have projected the data onto principal components (LSI) of the actor-actor co-variance matrix. The eigenvectors of this matrix are called 'eigencasts' and represent characteristic communities of actors that tend to co-appear in movies. The sparsity and magnitude of the network means that the components are dominated by communities with very small intersections, however, a closer look at such scatter plots reveals detail suggesting that a simple linear mixture model indeed provides a reasonable representation of the (small) coupling between these relative trivial disjunct subsets, see Figure 2. Such insight may be used for com-

puter assisted navigation of collaborative, peer-to-peer networks, for example in the context of search and retrieval.

Musical genre. The growing market for digital music and intelligent music services creates an increasing interest in modeling of music data. It is now feasible to estimate consensus musical genre by *supervised* learning from rather short music segments, say 5-10 seconds, see e.g., [18], thus enabling computerized handling of music request at a high cognitive complexity level. To understand the possibilities and limitations for unsupervised modeling of music data we here visualize a small music sample using the latent semantic analysis framework. The intended use is for a music search engine function, hence, we envision that

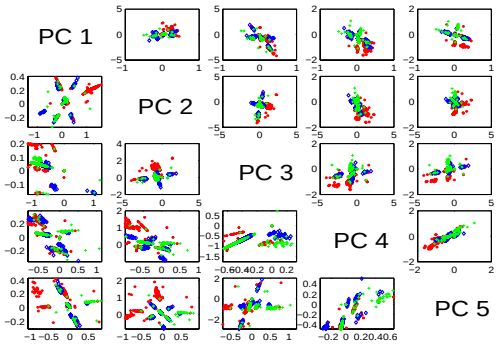

Fig. 3. We represent three music tunes (genre labels: `heavy metal, jazz, classical`) by their spectral content in overlapping small time frames ($w = 30$msec, with an overlap of 10msec, see [18], for details). To make the visualization relatively independent of 'pitch', we use the so-called mel-cepstral representation (MFCC, $K = 13$ coefficients pr. frame). To reduce noise in the visualization we have 'sparsified' the amplitudes. This was achieved simply by keeping coefficients that belonged to the upper 5% magnitude percentile. The total number of frames in the analysis was $F = 10^5$. Latent semantic analysis provided unsupervised subspaces with maximal variance for a given dimension. We show the scatter plots of the data of the first 1-5 latent dimensions. The scatter plots below the diagonal have been 'zoomed' to reveal more details of the ICA 'ray' structure. For interpretation we have coded the data points with signatures of the three genres involved: classical ($*$), heavy metal (diamond), jazz ($+$). The ICA ray structure is striking, however, note that the situation is not one-to-one (ray to genre) as in the small text databases. A component (ray) quantifies a characteristic musical 'theme' at the temporal level of a frame (30msec), i.e., an entity similar to the 'phoneme' in speech.

a largely text based query has resulted in a few music entries, and the algorithm is going to find the group structure inherent in the retrieval for the user. We

represent three tunes (with human genre labels: `heavy, jazz, classical`) by their spectral content in overlapping small time frames (w = 30msec, with an overlap of 10msec, see [18], for details). To make the visualization relatively independent of 'pitch', we use the so-called mel-cepstral representation (MFCC, $K = 13$ coefficients pr. frame). To reduce noise in the visualization we have further 'sparsified' the amplitudes. PCA provided unsupervised latent semantic dimensions and a scatter plot of the data on the subspace spanned by two such dimensions is shown in Figure 3. For interpretation we have coded the data points with signatures of the three genres involved. The ICA ray structure is striking, however, we note that the situation is not one-to-one as in the small text databases. A component quantifies a characteristic 'theme' at the temporal scale of a frame (30msec), it is an issue for further research whether genre *recognition* can be done from the salient themes, or we need to combine more than one theme to reach the classification performance obtained in [18].

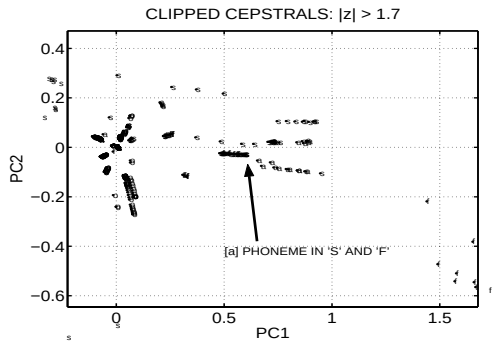

Fig. 4. Four simple utterances s, o, f, a were analysed. We analysed 40 msec windows of length (95% overlap). The windows were represented by 16 Mel-cepstrum coefficients. After variance normalization the features were sparsified based on energy zeroing windows of normalized magnitudes with a statistical $z \leq 1.7$. This threshold process retains 55% of the power in the original features. LSI/PCA was then performed on the sparsified feature coefficients for visualization. The results seem to indicate that generalizable cognitive components corresponding to the phoneme /ae/ opening the utterances s and f, can be identified using linear component analysis.

Phonemes as cognitive components of speech. There is a strong recent interest in representations and methods for computational auditory scene analysis, see e.g., Haykin and Chen's review on the cocktail party problem [19]. Low level cognitive components of speech encompass language specific features such as phonemes and speaker's voice prints. Such features can be considered 'pre-

Appendix E

Cognitive Components of Speech at Different Time Scales

This article is published in *Proc. 29th annual meeting of the Cognitive Science Society* 2007, pp 983-988, with the same title. Authors are Ling Feng and Lars Kai Hansen. It is also available as IMM publication database with number imm4871.

Cognitive Components of Speech at Different Time Scales

Ling Feng (lf@imm.dtu.dk)
Informatics and Mathematical Modelling
Technical University of Denmark
2800 Kgs. Lyngby, Denmark

Lars Kai Hansen (lkh@imm.dtu.dk)
Informatics and Mathematical Modelling
Technical University of Denmark
2800 Kgs. Lyngby, Denmark

Abstract

Cognitive component analysis (COCA) is defined as unsupervised grouping of data leading to a group structure well-aligned with that resulting from human cognitive activity. We focus here on speech at different time scales looking for possible hidden 'cognitive structure'. Statistical regularities have earlier been revealed at multiple time scales corresponding to: phoneme, gender, height and speaker identity. We here show that the same simple unsupervised learning algorithm can detect these cues. Our basic features are 25-dimensional short-time Mel-frequency weighted cepstral coefficients, assumed to model the basic representation of the human auditory system. The basic features are aggregated in time to obtain features at longer time scales. Simple energy based filtering is used to achieve a sparse representation. Our hypothesis is now basically ecological: We hypothesize that features that are essentially independent in a reasonable ensemble can be efficiently coded using a sparse independent component representation. The representations are indeed shown to be very similar between supervised learning (invoking cognitive activity) and unsupervised learning (statistical regularities), hence lending additional support to our cognitive component hypothesis.

Keywords: Cognitive component analysis; time scales; energy based sparsification; statistical regularity; unsupervised learning; supervised learning.

Introduction

The evolution of human cognition is an on-going interplay between statistical properties of the ecology, the process of natural selection, and learning. Robust statistical regularities will be exploited by an evolutionary optimized brain (Barlow, 1989). Statistical independence may be one such regularity, which would allow the system to take advantage of factorial codes of much lower complexity than those pertinent to the full joint distribution. In (Wagensberg, 2000), the success of given 'life forms' is linked to their ability to recognize independence between predictable and un-predictable process in a given niche. This represents a precision of the classical Darwinian paradigm by arguing that natural selection simply favors innovations which increase the independence of the agent and un-predictable processes. The agent can be an individual or a group. The resulting human cognitive system can model complex multi-agent scenery, and use a broad spectrum of cues for analyzing perceptual input and for identification of individual signal producing processes.

The optimized representations for low level perception are indeed based on independence in relevant natural ensemble statistics. This has been demonstrated by a variety of independent component analysis (ICA) algorithms, whose representations closely resemble those found in natural perceptual systems. Examples are, e.g., visual features (Bell & Sejnowski, 1997; Hoyer & Hyvrinen, 2000), and sound features (Lewicki, 2002).

Within an attempt to generalize these findings to higher cognitive functions we proposed and tested the independent cognitive component hypothesis, which basically asks the question: *Do humans also use information theoretically optimal ICA methods in more generic and abstract data analysis?* Cognitive component analysis (COCA) is thus simply defined as the process of unsupervised grouping of abstract data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity (Hansen, Ahrendt, & Larsen, 2005). For the preliminary research on COCA, human cognitive activity is restricted to the human labels in supervised learning methods. This interpretation is not comprehensive, however it is capable of representing some intrinsic mechanism of human cognition. Further more, COCA is not limited to one specific technique, but rather a conglomerate of different techniques. We envision that efficient representations of high level processes are based on sparse distributed codes and approximate independence, similar to what has been found for more basic perceptual processes. As mentioned, independence can dramatically reduce the perception-to-action mappings by using factorial codes rather than complex codes based on the full joint distribution. Hence, it is a natural starting point to look for high-level statistically independent features when aiming at high-level representations. In this paper we focus on cognitive processes in digital speech signals. The paper is organized as follows: First we discuss the specifics of the cognitive component hypothesis in relation to speech, then we describe our specific methods, present results obtained for the TIMIT database, and finally, we conclude and draw some perspectives.

Cognitive Component Analysis

In sensory coding it is proposed that visual system is near to optimal in representing natural scenes by invoking 'sparse distributed' coding (Field, 1994). The sparse signal consists of relatively few large magnitude samples in a background of numbers of small signals. When mixing such indepen-

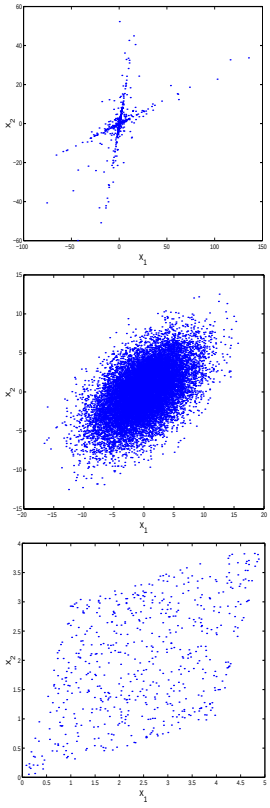

Figure 1: Prototypical feature distributions produced by a linear mixture, based on sparse (top), normal (middle), or dense source signals (bottom), respectively. The characteristics of a sparse signal is that it consists of relatively few large magnitude samples on a background of weak signals, hence, produces a characteristic ray structure in which the ray is defined by the vector of linear mixing coefficients: One for each for a sparse source.

dent sparse signals in a simple linear mixing process, we obtain the 'ray structure' which we consider emblematic for our approach, see the top panel in Figure 1. If a signal representation exists with a ray structure ICA can be used to recover both the line directions (mixing coefficients) and the original independent sources signals. Thus, we used ICA to model the ray structure and represent semantic structure in text, social networks, and other abstract data such as music

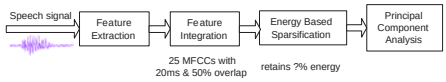

Figure 2: Preprocessing pipeline for speech COCA. MFCCs are extracted at the basic time scale (20*ms*). According to applications, features are averaged/stacked into longer time scales. Energy based sparsification is followed as a method to reduce intrinsic noise. PCA on sparsified features projects on a relevant subspace that makes it possible to visualize the 'ray'-structure. A subsequent ICA can be used to identify the actual ray coordinates and source signals.

(Hansen et al., 2005; Hansen & Feng, 2006). Within so-called bag-of-words representations of text, COCA is a generalization of principal component analysis based 'latent semantic analysis' (LSA), originally developed for information retrieval on text (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). *The key observation is that by using ICA, rather than PCA, we are not restricted to orthogonal basis vectors*. Hence, in ICA based latent semantic analysis topic vocabularies can have large overlaps. We envision that these implemented by overlapping receptive fields can detect more subtle differences than 'orthogonal' receptive fields.

Here we are going to elaborate on our earlier findings related to speech. The basic preprocessing pipeline for COCA of speech is shown in Figure 2. First, basic features are extracted from a digital speech signal leading to a fundamental representation that shares two basic aspects with the human auditory system: A logarithmic dependence on signal power and a simple bandwidth-to-center frequency scaling so that our frequency resolution is better at lower frequencies. These so-called mel-frequency cepstral coefficients[1] (MFCC) features are next aggregated in time. Simple energy based filtering leads to sparse representations. Sparsification is regarded as a simple means to emulate a saliency based attention process.

We have earlier reported our preliminary findings of ICA ray structure related to phonemes and speaker identity in a relatively small database (Feng & Hansen, 2005, 2006). Figure 3 illustrates the phoneme relevant ray structure at the basic time scale. This analysis was carried out on four simple utterances: 's', 'o', 'f' and 'a'. As shown in the figure, cognitive components of /e/ phoneme opening 's' and 'f' are identified.

We speculate that these phoneme-relevant cognitive components contribute towards the well-known basic invariant 'cue' characteristics of speech (Blumstein & Stevens, 1979). The theory of acoustic invariants points out that the perceived signals are derived as stable phonetic features despite of the different acoustic properties produced by different speakers. Moreover Damper has shown that although the speech signal may vary due to coarticulation, the relation between key fea-

[1] For a complete description of MFCC and related cepstral coefficients, see (Deller, Hansen, & Proakis, 2000).

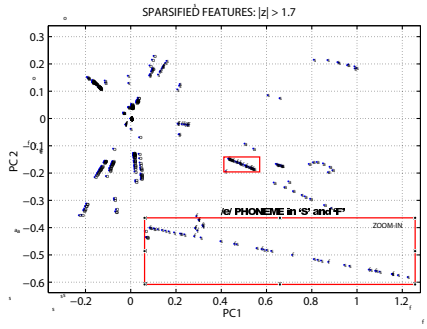

Figure 3: The latent space is formed by the two first principal components of data consisting of four separate utterances representing the sounds 's', 'o', 'f', 'a'. The structure clearly shows the sparse component mixture, with 'rays' emanating from the origin (0,0). The ray embraced in a rectangle contains a mixture of 's' and 'f' features, a cognitive component associated with the vowel /e/ sound.

tures follows a consistent and invariant form (Damper, 1998). Experiments involving labels related to speaker identification also provided the signature of linear 'ray'-structures. Is linearity related to perceptually distinguishable categories? The discussion on linear correlations in the speech signal and locus equation is still on-going (Sussman, Fruchter, Hillbert, & Sirosh, 1998).

During the itinerary of searching for spoken cognitive components, we have thus already reported (Feng & Hansen, 2005, 2006) on generalizable phoneme relevant components at a time scale of $20 \sim 40ms$, and generalizable speaker specific components at an intermediate time scale of $1000ms$.

In this paper we will further expand on our findings in speech by applying COCA on speech features at various time scales. We will systematically investigate the performance of unsupervised and supervised learning and test whether the tasks are learned in equivalent representations, hence, indicating consistency of statistical regularities (unsupervised learning) and human cognitive processes (supervised learning of human labels).

Methods

Our speech analysis follows the basic preprocessing scheme shown in Figure 2.

Feature Stacking

Since speech signals are non-stationary features have to be extracted from short-time scales. A simple method to get features at longer time scales is stacking or vector 'concatenation' of signals. Figure 4 illustrates the stacking procedure used in our experiments.

1. Truncate speech signal into overlapped frames, $20ms$ long with 50% overlap;

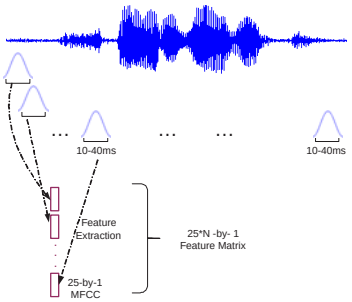

Figure 4: Speech feature extraction and stacking

2. Apply hamming window on each frame;
3. Extract MFCCs from each windowed frame, which forms a 25-dimensional vector;
4. According to the time scale, N original 25-dimensional MFCCs are stacked into one $25 * N$-dimensional vector;
5. Repeat 4 until all the frames are stacked.

$25 * N$ dimensional features representing long time scales are then used in both supervised and unsupervised learning methods.

Mixture of Factor Analyzers

To test whether supervised and unsupervised learning lead to similar representations we need a model that can incorporate both. In particular we need a generative representation to allow unsupervised learning, and we want the representation to

allow sparse linear ray like features. This can be achieved in a simple generalization of so-called mixture of factor analyzers (MFA). The unsupervised version is inspired by the so-called *Soft-LOST* (Line Orientation Separation Technique) (O'Grady & Pearlmutter, 2004).

Factor analysis is one of the basic dimensionality reduction forms. It models the covariance structure of multi-dimensional data by expressing the correlations in lower dimensional latent subspace, mathematical expression is

$$\mathbf{x} = \Lambda\mathbf{z} + \mathbf{u}, \tag{1}$$

where $\mathbf{x}$ is the p-dimensional observation; Λ is the factor loading matrix; $\mathbf{z}$ is the k-dimensional hidden factor vector which is assumed Gaussian distributed, $\mathcal{N}(\mathbf{z}|0,I)$; $\mathbf{u}$ is the independent noise which is $\mathcal{N}(\mathbf{u}|0,\Psi)$, with a diagonal matrix Ψ. Given eq. (1), observations are also distributed as $\mathcal{N}(\mathbf{x}|0,\Sigma)$, with $\Sigma = \Lambda\Lambda^T + \Psi$. Factor analysis aims at estimating Λ and Ψ in order to give a good approximation of covariance structure of $\mathbf{x}$.

While the simple factor analysis model is globally linear and Gaussian, we can model non-linear non-Gaussian processes by invoking a so-called mixture of factor analyzers

$$p(\mathbf{x}) = \sum_{i=1}^{K} \int p(\mathbf{x}|i,\mathbf{z})p(\mathbf{z}|i)p(i)d\mathbf{z}, \tag{2}$$

where $p(i)$ are mixing proportions and K is the number of factor analyzers. MFA combines factor analysis and the Gaussian mixture model, and hence can simultaneously perform clustering, and dimensionality reduction within each cluster, see (Ghahramani & Hinton, 1996) for a detailed review.

To meet our request for unsupervised learning model, MFA is modified to form an ICA-like line based density model similar to *Soft-LOST* by reducing the factor loadings to hold a single column vector, i.e., the 'ray' vector. It uses an EM procedure to identify orientations within a scatter plot: in the E-step, all observations are *soft* assigned into K clusters depending on the number of mixtures, which is represented by orientation vectors v_i, then it calculates posterior probabilities assigning data points to lines; and in M-step, covariance matrices are calculated for K clusters, and the principal eigenvectors of covariance matrices are used as new line orientations v_i^{new}, by this means it re-positions the lines to match the points assigned to them. Finally we end up with a mixture of lines which can be used as a classifier. We purposed a supervised mode of the modified MFA, which models the joint distribution of features set $\mathbf{x}$ and a possible labels set $\mathbf{y}$

$$p(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{K} \int p(\mathbf{x}|i,\mathbf{z})p(\mathbf{z})d\mathbf{z}p(\mathbf{y}|i)p(i). \tag{3}$$

In the sequel we will compare the performance of the two modes of modified MFA at multiple time scales. In particular we will train supervised and unsupervised models on the same feature set. For the unsupervised model we first train using only the features $\mathbf{x}$. When the density model is optimal we clamp the mixture density model and train only the cluster tables $p(\mathbf{y}|i)$, $i = 1,...,K$, using the training set labels. This is also referred to as unsupervised-then-supervised learning. This is a simple protocol for checking the cognitive consistency: Do we find the same representations when we train them with and without using 'human cognitive labels'.

Results

In this section we will present experimental results of analysis on speech signals gathered from TIMIT database (Garofolo et al., 1993). TIMIT is a reading speech corpus designed for the acquisition of acoustic-phonetic knowledge and for automatic speech recognition systems. It contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from the United States. For each utterance we have several labels that we think as cognitive indicators, labels that humans can infer given sufficient among of data. While each sentence lasts approximately $3s$ we will investigate performance at time scales ranging from basic $20ms$ to long about $1000ms$. The cognitive labels we will focus on here are phonemes, gender, height and speaker identity. Training and test sets are recommended in TIMIT, which contain 462 speakers reading for training and 168 for test. The total speech covers 59 phonemes, and the heights from all speakers range from $4'9''$ to $6'8''$, and have totally 22 different values. In order to gather sufficient amount of speech signals we chose 46 speakers with equal gender distribution, and speech signals cover all 59 phonemes, and all 22 heights.

Following the preprocessing pipeline, we first extracted 25-dimensional MFCCs from original digital speech signals. To investigate various time scales, we stacked basic features into a variety of time scales, from the basic $20ms$ scale up to $1100ms$. Energy based sparsification was used afterwards as a means to reduce the intrinsic noise and to obtain sparse signals. Sparsification is done by thresholding the amplitude of stacked MFCC coefficients, and only coefficients with super threshold energy were retained. By adjusting the threshold, we examine the role of sparsification in our experiments. We changed the threshold leading to a retained energy from 100% to 41%. Unsupervised and supervised modes of MFA were then performed respectively. To classify a new datum point x_{new} we first calculate the set of $p(i|\mathbf{x}_{new})$'s and then compute the posterior label probability.

Figure 5 presents the results of MFA for gender detection. The two plots (a) and (b) show the error rates for the supervised mode of MFA for the training and test set separately, while (c) and (d) are training and test error rates for unsupervised MFA (*soft-LOST*). First, we note that sparsification does play a role: when high percentage of features was retained from sparsification, e.g. 100% and 99.8%, error rates did not change much while increasing time scales, meaning the intrinsic noise covers up the informative part, and longer time scales do not assist to recover it. With the increasing of time scales all the curves tend to converge at the time scale around $400 \sim 500ms$.

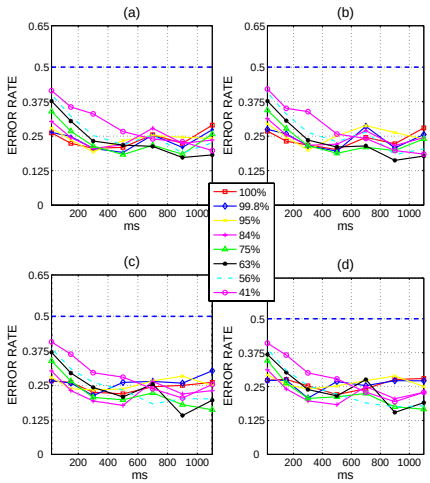

Figure 5: Error rates as function of time scales for different thresholds in gender detection. (a), (b): Training error rates and test error rates of supervised MFA respectively; (c), (d): Training error rates and test error rates of unsupervised MFA; The 8 curves represent feature sparsification with retained energy from 100% to 41%. The dashed lines are the baseline error rates for random guessing. Results indicate that the relevant time scale is about $400 \sim 500ms$ for this task.

Table 1: Timescales recommended for modeling Phonemes, Gender, Height, Identity

(*ms*)	**Phoneme**	**Gender**	**Height**	**ID**
Timescale	20	400-500	≥ 1000	≥ 1000

Similar experiments have been performed on phoneme, height and speaker identity. For phoneme recognition, the 59 phonemes from TIMIT database include vowels, fricatives, stops, affricates, nasals, semivowels and glides. To simplify the problem, we grouped these phonemes into 3 large categories: Vowels, fricatives and others. Stacking features into longer time scales for phoneme recognition degrades the performance, which shows consistency with our previous work that phonemes are best modeled at short time scale. The results of all experiments are summarized in Table 1.

To illustrate how well supervised and unsupervised representations are aligned, we follow the approach outlined above. We trained with appropriate labels in supervised mode to represent the human observer, and with the unsupervised-then-supervised scheme to represent the 'ecological' grouping. In both cases we can measure the test performance of the resulting classifier. High correlation between the error rates of the two schemes indicates similarity of the representations.

Figure 6 presents the correlation of test performance for supervised and unsupervised learning modes of MFA. For all the four classification tasks, for the given time scales and thresholds, data show a remarkable correlation. Hence, in line with the cognitive component hypothesis the statistical regularities captured by unsupervised learning are highly compatible with the cognitive structure represented by the label structures.

Conclusion

Cognitive component analysis of speech have revealed statistical regularities at multiple time scales corresponding to phoneme, gender, height and speaker identity.

We have devised a protocol for testing the cognitive component hypothesis based. We propose to compare the performance of supervised learning and unsupervised learning under closely matched conditions, so that the only difference is that 'cognitive labels' are used for supervised learning while not for unsupervised learning.

We preprocessed speech in a pipeline starting from the basic features: short time (20*ms*) 25-dimensional Mel-frequency Cepstral Coefficients (MFCCs). Feature stacking was used to aggregate features at multiple time scales. Energy based sparsification was invoked to obtain a sparse distributed representation and for noise reduction. We found that

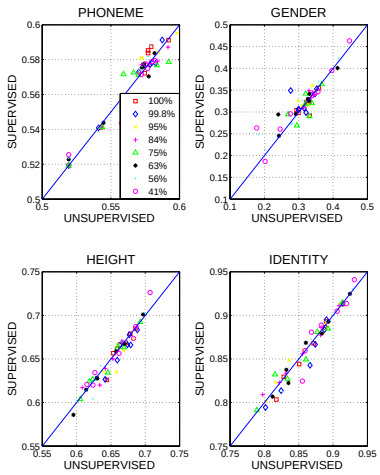

Figure 6: Correlation between test error rates of supervised and unsupervised learning on four label sets: phoneme, gender, height and identity. Solid lines indicate $y = x$ in the given coordinate systems. All data locate along this line. We can conclude that high correlation between supervised and unsupervised learning has been found for a wide variety of error rates substantiating our claim that two representations are highly similar.

the following time scales are characteristic: 20*ms* of speech provides phonemes information; gender is found in the range $400 \sim 500ms$; while, height and identity may require longer time scales, say $> 1000ms$.

Our finding indeed indicates the consistency of statistical regularities (unsupervised learning) and human cognitive processes (supervised learning of human labels), for phonemes, gender, speaker identity all of which are effortless recognized by humans. Height is also predicted from speech features corresponding to human ability to guess the speakers size. It would be interesting to test whether our representations lead to similar errors in predicting a persons height from speech as in humans.

Acknowledgments

This work is supported by the Danish Technical Research Council, through the framework project 'Intelligent Sound' (STVF No. 26-04-0092), www.intelligentsound.org. We thank Tobias Andersen for useful comments on the manuscript. LF thanks the Niels Bohr Legatet for generous financial support for external research stay.

References

Barlow, H. (1989). Unsupervised learning. *Neural Computation*, *1*, 295–311.

Bell, A. J., & Sejnowski, T. J. (1997). The 'independent components' of natural scenes are edge filters. *Vision Research*, *37*, 3327–3338.

Blumstein, S. E., & Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *The Journal of the Acoustical Society of America*, *66*, 1001–1017.

Damper, R. I. (1998). Self-learning and self-organization as tools for speech research. *Behavioral and brain sciences*, *21*, 262–263.

Deerwester, S. C., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*, 391–407.

Deller, J. R., Hansen, J. H., & Proakis, J. G. (2000). *Discrete time processing of speech signals*. IEEE Press Marketing.

Feng, L., & Hansen, L. K. (2005). On low level cognitive components of speech. In *Proc. international conference on computational intelligence for modelling* (Vol. 2, pp. 852–857).

Feng, L., & Hansen, L. K. (2006). Phonemes as short time cognitive components. In *Proc. icassp* (Vol. 5, p. 869-872).

Field, D. J. (1994). What is the goal of sensory coding? *Neural Computation*, *6*, 559–601.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., & Dahlgren, N. L. (1993). The darpa timit acoustic phonetic continuous speech corpus cdrom. In *Nist order number pb91-100354*.

Ghahramani, Z., & Hinton, G. E. (1996). *The em algorithm for mixtures of factor analyzers* (Tech. Rep. No. CRG-TR-96-1). 6 King's College Road, Toronto, Canada M5S 1A4: University of Toronto, Department of Computer Science.

Hansen, L. K., Ahrendt, P., & Larsen, J. (2005). Towards cognitive component analysis. In *Akrr'05 -international and interdisciplinary conference on adaptive knowledge representation and reasoning*.

Hansen, L. K., & Feng, L. (2006). Cogito componentiter ergo sum. In *Proc. ica* (pp. 446–453).

Hoyer, P., & Hyvrinen, A. (2000). Independent component analysis applied to feature extraction from colour and stereo images. *Network: Comput. Neural Syst.*, *11*, 191–210.

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, *5*, 356–363.

O'Grady, P. D., & Pearlmutter, B. A. (2004). Soft-lost: Em on a mixture of oriented lines. In *Proc. ica* (p. 430-436).

Sussman, H. M., Fruchter, D., Hillbert, J., & Sirosh, J. (1998). Linear correlations in the speech signal: The orderly output constraint. *Behavioral and brain sciences*, *21*, 241–299.

Wagensberg, J. (2000). Complexity versus uncertainty: The question of staying alife. *Biology and philosophy*, *15*, 493–508.

Appendix F

On Phonemes as Cognitive Components of Speech

This article is accepted for publication in *Proc. IAPR Workshop on Cognitive Information Processing* 2008, pp 205-210, with the same title. Authors are Ling Feng and Lars Kai Hansen. It is also available as IMM publication database with number imm5609.

ON PHONEMES AS COGNITIVE COMPONENTS OF SPEECH

Ling Feng, Lars Kai Hansen

Technical University of Denmark
Informatics and Mathematical Modelling
Richard Petersens Plads, B 321

ABSTRACT

COgnitive Component Analysis (COCA) defined as the process of unsupervised grouping of data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity, has been explored on phoneme data. Statistical regularities have been revealed at multiple time scales. The basic features are 25-dimensional short time ($20ms$) mel-frequency weighted cepstral coefficients. Features are integrated by means of stacking to obtain features at longer time scales. Energy based sparsification is carried out to achieve sparse representations. Our hypothesis is ecological: we assume that features that essentially independent in a context defined ensemble can be efficiently coded using a sparse independent component representation. This means that supervised and unsupervised learning should result in similar representations. We indeed find that supervised and unsupervised learning seem to identify similar representations, here, measured by the classification similarity.

Index Terms— Cognitive Component Analysis, Unsupervised Learning, Supervised Learning, Phoneme Classification.

1. INTRODUCTION

Cognition generally refers to capabilities of human minds, such as reasoning, perception, intelligence and learning, etc. The human cognitive system can model complex multi-agent scenery, and uses a broad spectrum of cues for analyzing perceptual input and for identification of individual signal process components. The purpose is to infer the proper action for a given situation. Robust statistical regularities can be exploited by an evolutionary optimized brain in making inference about appropriate actions [1]. *Statistical independence* is likely to be such regularity. Knowledge about an independence rule will allow the system to take advantage of a corresponding factorial code typically of (much) lower complexity than the one pertinent to the full joint distribution. The optimized representations of the low level cognition (perception) are known to be based on independence in the relevant natural ensemble statistics [2, 3]. This has led to a surge of interest in independent component analysis (ICA) for modeling perceptive tasks, and the resulting representations share many features with those found in natural perceptual systems. Examples are, e.g., in visual features [2, 3], and sound features [4].

Within an attempt to generalize these findings to a higher cognitive function, we proposed the cognitive component hypothesis which basically runs: *Human cognition uses theoretically optimal ICA-like representations for generic data analysis*. COgnitive Component Analysis (COCA) is wherefore defined as the process of unsupervised grouping of generic data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity, see e.g., [5, 6]. In sensory coding it is proposed that visual system is near optimal in representing natural scenes by invoking 'sparse distributed' coding [7]. The sparse signal consists of relatively few large magnitude samples in a background of numbers of small signals. We envision that auditory areas of the perceptual system also abide by the sparse coding rule. When mixing such independent sparse signals in a simple linear mixing process, we obtain the 'ray structure' emblematic for cognitive component analysis. If a signal representation exists with a ray structure, ICA can be used to recover both the line directions (mixing coefficients) and the original independent source signals. Figure 1 illustrates the ray-structure representation of phoneme classification within three classes: vowels, fricatives, and stops.

Thus far, ICA has been used to model the ray structure and to represent semantic structure in text, social networks, and other abstract data, e.g. music [5, 8] and speech [9].

Since the mechanisms of human cognitive activity are still not fully understood, to quantify cognition may seem ambiguous and may also be considered way too ambitious. However, the direct consequence of cognition, human behavior, has a rich phenomenology that can be accessed and modeled. In the following analysis, we represent human cognition simply by a classification rule, i.e. based on a set of manually obtained labels we train a classifier using supervised learning. The question is then reduced to looking for similarity between the representations in supervised learning (of human labels) and unsupervised learning that simply explores the statistical properties of the domain. If a high correlation exists between the representations resulting from unsupervised and supervised

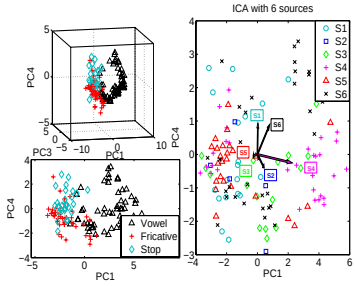

Fig. 1. Phoneme ray-structure. Figures on the left-hand side are scatter plots of phoneme features in the space of principal components. Data are displayed in different shapes denoting three classes: Vowels, Fricatives and Stops. Loosely speaking, fricatives and stops locate along solo-ray; and vowels spread more widely and can be represented by multi-rays. The right-hand side figure gives 6 independent sources. Arrows show the column vectors of the mixing matrix. By majority voting, source $1, 2$ stand for fricatives; $3, 4, 6$ for vowels; 5 for stops.

learning, we interpret this as the evidence that human cognition is based on the given statistical regularity. In this paper we will present a detailed comparison between unsupervised and supervised learning representations: at the classification rate level; at the sample-to-sample basis; and at the more detailed sample-to-sample posterior probability level. This paper focuses on cognitive component analysis of short time speech signals, to test whether phonemes are such cognitive components. First we discuss the preprocessing pipeline of COCA; secondly we introduce the unsupervised and supervised learning models; thirdly we systematically investigate the performance of unsupervised and supervised learning on the potential cognitive indicator: phoneme, and test whether the task is learnt in equivalent representations; and the conclusion summarizes this paper.

2. PREPROCESSING OF COGNITIVE COMPONENT ANALYSIS

Here we are going to elaborate on the speech-relevant cognitive component analysis. The basic preprocessing pipeline for speech COCA is shown in figure 2.

A efficient way of representing speech for machine speech analysis is usually to use spectral features of fairly low dimensionality, e.g. $20 \sim 30$ dimensions. The ideal features will be the ones which are capable of accounting for the functionality of human hearing system. The basic features in COCA analysis are extracted from digital speech signals leading to a fundamental representation that shares two basic aspects with the

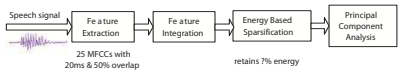

Fig. 2. Preprocessing pipeline for speech COCA. MFCCs are extracted at the basic time scale ($20ms$). Depending on the application features are integrated into longer time scales. Energy based sparsification is applied as a method to reduce intrinsic noise and get sparse representations. PCA projects features onto a base of cognitive processes. A subsequent ICA can be used to identify the actual ray coordinates and source signals.

human auditory system: A logarithmic dependence on signal power and a simple bandwidth-to-center frequency scaling so that the frequency resolution is better at lower frequencies. These so-called mel-frequency cepstral coefficients (MFCCs) can loosely represent the human auditory system response, which is triggered by the **mechanoreceptors** [10] of the inner ear, except that MFCCs can not model the outer ear which is critical for sound localization and loudness accuracy. The vibrations caused by the sound pressure waves receiving at the outer ear deflect the hairlike cells in the inner ear and trigger nerve impulses.

The computation of MFCCs is based on the time-frequency analysis. Since speech signals are non-stationary, features must be extracted from short time intervals, i.e. $10 \sim 40$ ms. The fast fourier transform (FFT) transforms the convolution relation between the excitation sequence and the vocal system impulse response into production; and the logarithm, afterwards, provides us with the linear combination (addition between these two). The mel-frequency warping step changes the frequency scale from linear to mel-scale, which is approximately linear below $1k$Hz and logarithmic above. Finally discrete cosine transform (DCT) brings us to the mel-cepstrum. For detailed description, see [11].

2.1. Feature Stacking

For a feature to reveal the semantic meaning of an audio signal, analysis over a much longer period is necessary, usually from one second to several tens seconds [12]. Feature integration is by and large a way to combine the information from several short-time features into a long-term feature. A simple integration is the stacking, in other words, vector 'concatenation' of signals.

1. Truncate speech signals into short time frames, $20ms$ long with 50% overlap;
2. Apply hamming window on each frame;
3. Extract MFCCs from each frame, which forms (e.g.) a 25-dimensional vector;
4. According to the time scale, the MFCCs from the first N frames are stacked into one $25 * N$-d vector;

5. Repeat 4 with the next N short time frames (without overlap) until all the short time frames are stacked (and exclude the residual).

The resulting $25 * N$-d features representing long time scales are then further processed.

2.2. Energy Based Sparsification (EBS)

Simple energy based filtering leads to sparse representations. Sparsification is regarded as a simple means to filter out the small signals, which emulates a saliency based *attention* process related to **detectability** and **sensory magnitude** from perceptual principles [10]. For auditory perception only the signals reaching the postsynaptic cell's threshold will lead to the cell firing [13]. Therefore sparsification is done by thresholding the stacked features, and only coefficients with superior energy are retained, and the rest is set zero.

2.3. Principal Component Analysis

PCA is an orthogonal linear transformation technique. It is often used for dimensionality reduction, and in the meanwhile remains the most variance of the data. In textual information analysis PCA is known as LSA. It presumes that the semantic content of the overall document can be approximated as the word usage. The low-dimensional space transformed by PCA/LSA from high-dimensional space is regarded as the basis for all cognitive processing [14]. LSA has human-like performance in text analysis, we assume that it can also be used to get the relevant basis for speech cognitive related tasks. It has been proved that in some cases, LSA can provide good simulations of human cognitive processes alone, and in other cases it is often operated as base for cognitive processes.

Singular value decomposition (SVD) is invoked to identify a relevant signal subspace based simply on signal variance,

$$\mathbf{X} = \mathbf{U}\Lambda\mathbf{V}^T, \quad \mathbf{Y} = \mathbf{U}_k^T\mathbf{X}, \tag{1}$$

where $\mathbf{X}$ is a m-by-n data matrix, $\mathbf{U}$ is a m-by-m orthonormal matrix, Λ is a m-by-n matrix with the singular values along the diagonal, and $\mathbf{V}$ is a n-by-n orthonormal matrix. The dimensionality of data is reduced by projecting the data to the first k principal components ($k < m$).

3. MODELS

Having the comparison of the unsupervised and supervised learning in mind, we need to have two models which share similarities w.r.t the model structure. Moreover both models should allow sparse linear ray-like features. The Bayesian classifier which assumes a known probabilistic density distribution for each class, has been widely used and is misclassification error rate optimal. Here we choose two Bayesian classifiers: Naive Bayes and Mixture of Gaussians (MoG). For the unsupervised learning model we first apply unsupervised ICA only on the features. After recovering the source signals, we add the label information to a naive Bayes classifier, which assumes that the distribution of the source within each class is Gaussian. To keep the consistency of using Bayesian classifier and Gaussian model, we choose Mixture of Gaussians as the supervised learning model. This is a simple protocol for checking the cognitive consistency: Do we find the same representations when we train them with and without using 'human cognitive labels'?

3.1. Unsupervised Learning

As mentioned, if the sparse features are essentially independent, ICA can be used to recover both the mixing coefficients and the original independent sources. The typical algorithms for ICA use centering, whitening and dimensionality reduction as preprocessing steps in order to reduce the complexity of the algorithm. PCA is normally used to achieve the whitening and dimension reduction. Since in the preprocessing pipeline we have applied PCA on stacked and sparsified MFCC features, we directly apply ICA algorithm on PCA coefficients without dimensionality reduction.

The generative formula of noise free ICA model is

$$\mathbf{Y} = \mathbf{AS}, \tag{2}$$

where $\mathbf{Y}$ is the k-dimensional observation; $\mathbf{A}$ is the mixing matrix with dimension k-by-p; $\mathbf{S}$ is the matrix of p independent sources which are assumed non-Gaussian. ICA aims at estimating both the mixing matrix $\mathbf{A}$ and the sources $\mathbf{S}$. This is done by either maximizing the non-Gaussianity of the calculated sources or minimizing the mutual information.

The original sources can be recovered by

$$\mathbf{S} = \mathbf{WY}, \tag{3}$$

where we assume the total no. of sources (k) is the same as the dimension of the observation $\mathbf{y}$ (p) in the following experiments, hereby $\mathbf{W} = \mathbf{A}^{-1}$ is the unmixing matrix, and the $\mathbf{A}$ and $\mathbf{W}$ matrices are therefore square.

To reveal the performance of unsupervised learning in classification tasks, we first train the unsupervised model using only the features (principal components) $\mathbf{Y}$ to recover the sources $\mathbf{S}$. Since sources are independent, then naive Bayes classifier can be applied on sources with the training set labels. This is also referred to as unsupervised -then- supervised learning scheme.

The naive Bayes classifier assumes independency of input feature for each class, and is based on Bayes' theorem:

$$p(\mathbf{C}_i|\mathbf{s}) = \frac{p(\mathbf{s}|\mathbf{C}_i)p(\mathbf{C}_i)}{\sum_i p(\mathbf{s}|\mathbf{C}_i)p(\mathbf{C}_i)} \tag{4}$$

where $p(\mathbf{C}_i)$ denotes the i^{th} class prior; $p(\mathbf{s}|\mathbf{C}_i)$ is the likelihood of the $\mathbf{C}_i$; and $p(\mathbf{C}_i|\mathbf{s})$ is the posterior of the i^{th} class given data $\mathbf{s}$: $\mathbf{s} = (s_1, \ldots, s_p)^T$.

As naive Bayes assumes that the data input variables are independent, the likelihood in equation (4) can be simplified as:

$$p(\mathbf{s}|\mathbf{C}_i) = \prod_{n=1}^{p} p(s_n|\mathbf{C}_i), \tag{5}$$

where each $p(s_n|\mathbf{C}_i)$ is modeled as univariate Gaussian distribution $\mathcal{N}(\mu_{ni}, \sigma_{ni}^2)$.

For the classification problem, we apply the $\mathbf{W}$ learnt from training set to new data $\mathbf{Y}^{new}$, and recover their sources $\mathbf{S}^{new}$. Afterwards, the trained naive Bayes classifier with a set of Gaussian parameters (means and variances) will be used on $\mathbf{S}^{new}$ to predict the labels of new data.

3.2. Supervised Learning

As for the supervised learning model, we intend to choose a very flexible model, which is able to represent human decisions. We here use the Mixture of Gaussians,

$$p(\mathbf{C}_i|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{C}_i)p(\mathbf{C}_i)}{\sum_i p(\mathbf{y}|\mathbf{C}_i)p(\mathbf{C}_i)}, \tag{6}$$

and the likelihood will be,

$$p(\mathbf{y}|\mathbf{C}_i) = \sum_j p(\mathbf{y}|j, \mathbf{C}_i)p(j|\mathbf{C}_i), \tag{7}$$

where $p(\mathbf{y}|j, \mathbf{C}_i) = \mathcal{N}(\mathbf{y}|\mathbf{m}_{ji}, \mathbf{V}_{ji})$, and $p(j|\mathbf{C}_i)$ is the mixing parameters in class $\mathbf{C}_i$. The parameters $\mathbf{m}_{ji}$, $\mathbf{V}_{ji}$ are estimated from the training set via the standard Expectation-Maximization algorithm. For simplicity, we assume the covariance matrices to be diagonal. Note that although features are independent within each mixture component due to the diagonal covariance matrix, the mixture model does not factorize over features. The MoG is capable of modeling arbitrary dependency structures among features [15] if the number of mixture components is sufficiently large. On the other hand, a MoG with many mixture components is prone to overfitting and will most likely not generalize well. In our experiments, we vary the number of mixture components, and select models according to classification accuracy. Observations are assigned to the class having the maximum *posterior* probability. Maximum A *Posteriori* (MAP) criterion aims at maximizing the *posterior* $p(\mathbf{C}|\mathbf{y})$ rather than the likelihood $p(\mathbf{y}|\mathbf{C})$.

4. EXPERIMENTAL DESIGN AND RESULTS

4.1. Experimental Design

The experiments were carried out on speech signals gathered from TIMIT database [16]. TIMIT collects reading speech from 630 native English speakers. There are totally 10 sentences from individual speaker, while each lasts approximately $3s$. Here we focused on phoneme classification. Each sentence has been manually labeled with phonetic symbols. There are 60 phonemes in total. In order to gather a sufficient amount of speech, we chose 46 speakers with equal gender partition, and speech signals covered all 60 phonemes, including vowels, fricatives, stops, affricates, nasals, semivowels and glides. To simplify the classification problem, we pre-grouped them into 3 large categories: vowels, fricatives and others. The unsupervised and supervised models were compared in a set of experiments: we stacked the basic time scale features into several longer time scales, and sparsified the stacked features with different degrees to test the consistency of the comparison. In the meanwhile of the performance comparison, we also anticipated to find out the role of time scales.

Following the preprocessing pipeline, we first extracted 25-d MFCCs from original speech signals with hamming windows in the time domain and triangular filters in the mel-frequency domain. Within these 25 dimensions, the so-called 0^{th} order MFCC was also included, which represents the log-energy of each short time frame. To investigate the role of time scales, we stacked the basic features into a variety of time scales, from $20ms$ scale up to $1100ms$ (20, 100, 150, 300, 500, 700, 900 and $1100ms$). Energy based sparsification was used afterwards. The degree of sparsification was controlled by thresholds leading to the retained energy from 100% to 65%. PCA was then carried out on stacked and sparsified features, and dimensionality of the features was reduced. For features having longer time scales than $20\ ms$, their dimensions were reduced to 100, and the dimension of the features at the basic time scale remained the same, i.e. 25.

After the preprocessing of features, we input the data into unsupervised and supervised models respectively. The training set covered 6 sentences from each of the 46 speakers, and the rest 4 sentences were used as test set. The ICA algorithm evaluated the unmixing matrix $\mathbf{W}$ of the training set, and the sources $\mathbf{S}^{train}$ were consequently recovered in unsupervised learning. Afterwards the sources were input to the naive Bayes classifier together with training set labels to estimate the parameters of the independent univariate Gaussians. For prediction, we preprocessed the test set following the same procedure. The $\mathbf{W}$ derived from the training set was applied to the test set to recover the sources $\mathbf{S}^{test}$. Whereafter naive Bayes classifier predicted the labels of the test set based on the test sources. We have used the exact same training and test set for the supervised model as for the unsupervised model, so as to exclude the comparison bias introduced by data. MoG models estimated a set of Gaussian distributions for each class from the training set, and fulfilled the label prediction on the test set. Both models provided us with a set of labels and a set of posterior label probabilities for both data sets.

4.2. Results

A set of experiments were carried out in 64 (8 times 8) different conditions, i.e. 8 time scales and 8 sparsification levels.

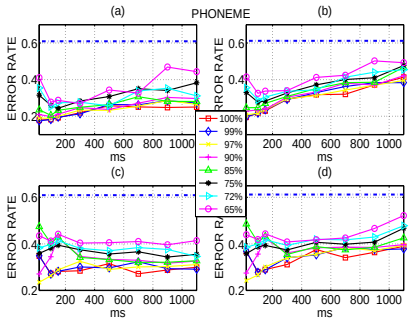

Fig. 3. Error rates as a function of time scales for different thresholds in phoneme classification. (a), (b): Training and test error rates of supervised MoG; (c), (d): Training and test error rates of unsupervised model, respectively; The 8 curves represent feature sparsification with retained energy from 100% to 65%. The dashed lines are the baseline error rates for random guessing. Results indicate that the relevant time scale is around the basic time scale.

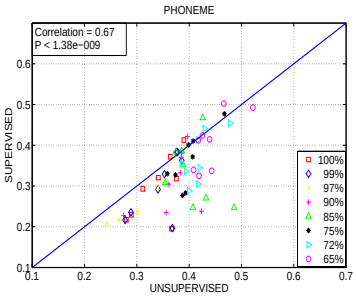

Fig. 4. Correlation between test error rates of supervised and unsupervised learning models. Solid lines indicate $y = x$. The correlation coefficient is 0.67.

Figure 3 presents the results of both supervised and unsupervised learning. The two plots (a) and (b) show the training and test error rates of the MoG models separately, whereas (c) and (d) are the training and test error rates of unsupervised learning (ICA+naive Bayes). The 8 curves in each panel represent the 8 EBS levels. First, it is quite obvious that features at longer time scales degraded the performance, which coincides with the conclusion from our previous research that phonemes are best modeled at short time scales [9, 17]. As we noticed, especially when retaining energy is 65%, high degree of sparsification decreased classification accuracy.

Error Rate Comparison From the above experiments we noticed that the performances of unsupervised and supervised models bear similarity w.r.t recognition error rates. To exam how well their representations are correlated, we measured the test performance of the resulting classifiers. High correlation between the error rates of the two schemes indicated similarity of the representations, shown in figure 4. The correlation is distinguished in phoneme classification task: for the given time scales and thresholds, data locate around $y = x$, and the correlation coefficient $\rho = 0.67, p < 1.38e - 009$.

Sample-to-Sample Correlation In order to reconfirm the finding and to account for the patterns of making decisions for both models, we followed the approach outlined above. We trained with the appropriate manual labels in supervised model to represent the human observer, and with the unsupervised -then- supervised learning scheme to represent the 'ecological' grouping. This experiment was also carried out on three groups of phonemes: vowels eh, ow; fricatives s, z, f, v; and stops k, g, p, t, where eh stands for the vowel in the word 'BET', and ow for the vowel in 'BOAT'. Figure 5 presents the sample-to-sample classification results of both models. 25-d MFCCs were first sparsified, to keep 99% energy, and then PCA reduced the dimension to 6, and the resulting features were modeled by unsupervised and supervised learning methods separately. It is clear that two models had a similar pattern of making the correct prediction and making mistakes, and the percentage of matching (correct predictions from both models and misclassified samples from both models) between supervised and unsupervised learning was up to 91%.

Posterior Probability Comparison So far we have seen that there is a close correspondence at the level of error rates and sample-to-sample classification. A more detailed comparison can be obtained by considering the posterior probabilities obtained on a sample basis. We chose one experiment of the phoneme classification ($100ms$ time scale with 97% remaining energy) among the 64 experiments mentioned above. Figure 6 presents the posterior probability comparison of fricatives models. If two models are the exact match, we should expect that the posterior probabilities locate along the diagonal of the histograms with high distribution at $(1, 1)$ and $(0, 0)$. The matching in this case is around 57%.

5. CONCLUSION

With the purpose of understanding the exploitation of statistical regularities in human cognitive activity, we investigated the Cognitive Component Analysis. We have devised a protocol for testing the cognitive component hypothesis, that is to compare the performance of unsupervised learning, which aims at discovering statistical regularities, and supervised learning, which loosely represents human cognitive activity.

We have studies the COCA on phoneme level signals, and

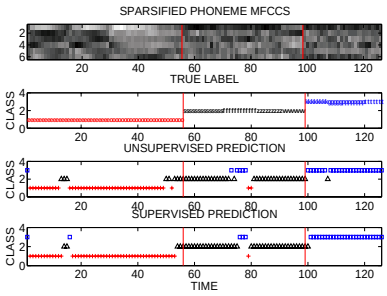

Fig. 5. Sample-to-sample phoneme classification among vowels, fricatives and stops. This first panel shows the temporal development of sparsified MFCCs. The boundaries of 3 phoneme classes are highlighted by vertical lines. Second panel gives the true labels, denoted by phonetic symbols. The last two panels give the unsupervised and supervised label predictions, marked by 3 shapes. The decision patterns of supervised and unsupervised learning show high similarity.

compared the performance of unsupervised and supervised learning at three levels: error rate level; sample-to-sample level; and the more detailed posterior probability level. In all the comparisons we have found evidence that supervised and unsupervised learning in fact do lead to similar representations.

6. ACKNOWLEDGEMENT

This work is supported by the Danish Technical Research Council, through the framework project 'Intelligent Sound' (STVF No. 26-04-0092), www.intelligentsound.org.

7. REFERENCES

[1] H.B. Barlow, "Unsupervised learning," *Neural Computation*, vol. 1, pp. 295–311, 1989.

[2] A. J. Bell and T. J. Sejnowski, "The 'independent components' of natural scenes are edge filters," *Vision Research*, vol. 37, pp. 3327–3338, 1997.

[3] P. Hoyer and A. Hyvrinen, "Independent component analysis applied to feature extraction from colour and stereo images," *Network: Comput. Neural Syst.*, vol. 11, pp. 191–210, 2000.

[4] M. S. Lewicki, "Efficient coding of natural sounds," *Nature Neuroscience*, vol. 5, pp. 356–363, 2002.

[5] L. K. Hansen, P. Ahrendt, and J. Larsen, "Towards cognitive component analysis," in *AKRR'05 -International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, 2005.

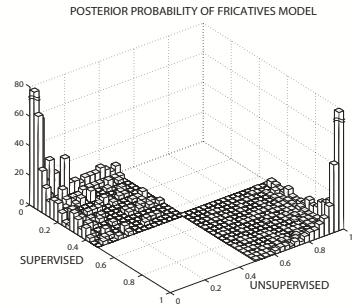

Fig. 6. Posterior probability comparison. Figure shows the histograms of the posterior probabilities provided by unsupervised and supervised fricatives models on the test set in the matching case. The two highest distributions locate at $(1, 1)$ and $(0, 0)$, which are 840 and 501 respectively.

[6] L. Feng and L. K. Hansen, "On low level cognitive components of speech," in *Proc. International Conference on Computational Intelligence for Modelling*, 2005, vol. 2, pp. 852–857.

[7] D. J. Field, "What is the goal of sensory coding?," *Neural Computation*, vol. 6, pp. 559–601, 1994.

[8] L. K. Hansen and L. Feng, "Cogito componentiter ergo sum," in *Proc. ICA*, 2006, pp. 446–453.

[9] L. Feng and L. K. Hansen, "Phonemes as short time cognitive components," in *Proc. ICASSP*, 2006, vol. 5, pp. 869–872.

[10] G. Mather, *Foundations of Perception*, Psychology Press, Hove, UK, 2006.

[11] J. R. Deller, J. H. Hansen, and J. G. Proakis, *Discrete Time Processing of Speech Signals*, IEEE Press Marketing, 2000.

[12] Y. Wang, Z. Liu, and J.C. Huang, "Multimedia content analysis using both audio and visual clues," *IEEE Signal Processing Magazine*, vol. 17, 2000.

[13] D. Reisberg, *Cognition: Exploring the Science of the Mind*, W.W.Norton & Company, New York, USA, 2006.

[14] W. Kintsch, "Predication," *Cognitive Science*, vol. 25, pp. 173–202, 2001.

[15] C. M. Bishop, *Neural Networks for Pattern Recognition*, OXFORD University Press, Oxford, UK, 1995.

[16] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett and N. L. Dahlgren, "The DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," in *NIST order number PB91-100354*, 1993.

[17] L. Feng and L. K. Hansen, "Cognitive components of speech at different time scales ," in *Proc. CogSci*, 2007, pp. 983–988.

Appendix G

Is Cognitive Activity of Speech Based on Statistical Independence?

This article is accepted for publication in *Proc. 30th annual meeting of the Cognitive Science Society* 2008, pp 1197-1202, with the same title. Authors are Ling Feng and Lars Kai Hansen. It is also available as IMM publication database with number imm5651.

Is Cognitive Activity of Speech Based on Statistical Independence?

Ling Feng (LF@Imm.Dtu.Dk)
Informatics and Mathematical Modeling
Technical University of Denmark

Lars Kai Hansen (LKH@Imm.Dtu.Dk)
Informatics and Mathematical Modeling
Technical University of Denmark

Abstract

This paper explores the generality of COgnitive Component Analysis (COCA), which is defined as the process of unsupervised grouping of data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity. The hypothesis of COCA is ecological: the essentially independent features in a context defined ensemble can be efficiently coded using a sparse independent component representation. Our devised protocol aims at comparing the performance of supervised learning (invoking cognitive activity) and unsupervised learning (statistical regularities) based on similar representations, and the only difference lies in the human inferred labels. Inspired by the previous research on COCA, we introduce a new pair of models, which directly employ the independent hypothesis. Statistical regularities are revealed at multiple time scales on phoneme, gender, age and speaker identity derived from speech signals. We indeed find that the supervised and unsupervised learning provide similar representations measured by the classification similarity at different levels.

Keywords: Cognitive component analysis; statistical regularity; unsupervised learning; supervised learning; classification.

Introduction

The human cognitive system models complex multi-agent scenery, e.g. perceptual input and individual signal process components, so as to infer the proper action for a given situation. While making inference of appropriate actions, an evolutionary brain is capable of exploiting the robust statistical regularities (Barlow, 1989). *Statistical independence* is a potential candidate of such regularities, which determine the characteristics of human cognition. The knowledge about an independence rule will allow the system to take advantage of a corresponding factorial code typically of (much) lower complexity than the one pertinent to the full joint distribution. The series exploration of the independence in the relevant natural ensemble statistics (Bell & Sejnowski, 1997; Hoyer & Hyvrinen, 2000; Lewicki, 2002) has led to a surge of interest in independent component analysis (ICA) for modeling perceptive tasks, and the resulting representations share many features with those found in natural perceptual systems. The cognitive component hypothesis, consequently, has been proposed which basically runs: *Human cognition uses information theoretically optimal ICA methods in generic and abstract data analysis*. The hypothesis is ecological: we assume that essentially independent features in a context defined ensemble can be efficiently coded using a sparse independent component representation. Built upon this base, COgnitive Component Analysis (COCA) was wherefore defined as the process of unsupervised grouping of generic data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity, see (Hansen, Ahrendt, & Larsen, 2005; Feng & Hansen, 2005).

'Sparse distributed' sensory coding is near optimal to represent natural scenes in visual system (Field, 1994). We envision that auditory areas of the perceptual system also abide by the sparse coding rule. A sparse signal consists of relatively few large magnitude samples in a background of numbers of small signals. The emblematic phenomenon of COCA, namely the 'ray structure', will be revealed if such independent sparse signals are mixed in a linear manner. At this point, ICA is able to recover both the line directions (mixing coefficients) and the original independent sources. Thus far, ICA has been used to model the ray structure and to represent the semantic structure in text, the communities in social networks, and other abstract data, e.g. music (Hansen et al., 2005; Hansen & Feng, 2006) and speech (Feng & Hansen, 2006). Figure 1 illustrates the ray-structure representation of a phoneme classification within three classes.

Since the mechanisms of human cognitive activity are not yet fully understood, to quantify cognition may seem ambiguous. Nevertheless, the direct consequence of cognition, human behavior, has a rich phenomenology that can be accessed and modeled. In the following analysis, we represent human cognition simply by a classification rule, i.e. based on a set of manually obtained labels we train a classifier using supervised learning. The question is then reduced to looking for similarities between the representations in supervised learning (of human labels) and unsupervised learning that simply explores the statistical properties of the domain. The high correlation between the representations resulting from unsupervised and supervised learning can be interpreted as the evidence that human cognition is based on the given statistical regularity.

Feng and Hansen (2007) have explored speech cognitive components at different time scales, and have shown that unsupervised and supervised learning based on modified mixture of factor analyzers (MFA) could identify similar representations. MFA has been modified to ICA-like line based density model. In this paper we will carry on the analysis of speech signals, and introduce a new pair of unsupervised and supervised models, where the unsupervised model directly reflects the independent hypothesis. Detailed comparisons between unsupervised learning of statistical properties and su-

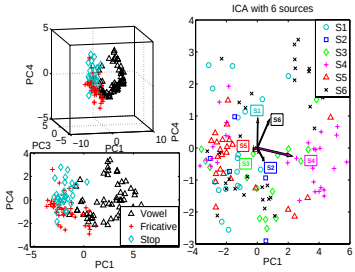

Figure 1: Phoneme ray-structure. The left-hand side panels are scatter plots of phoneme features in the space of principal components. Different shapes denote three classes: Vowels, Fricatives and Stops. The right-hand side panel gives 6 independent sources. The arrows show the column vectors of the mixing matrix. Loosely speaking, source $1,2$ stand for fricatives; $3,4,6$ for vowels; 5 for stops by majority voting.

pervised learning of human labels will be presented: at the classification rate level; at the sample-to-sample base; and at the more detailed sample-to-sample posterior probability level. Here COCA focuses on four potential cognitive indicators: phoneme, gender, age and identity.

Preprocessing of COCA

The basic preprocessing pipeline for COCA analysis of speech is given in Figure 2.

To use spectral features of fairly low dimensionality, e.g. $20 \sim 30$, is a common way to represent speech for machine analysis. The ideal features are expected to be capable of accounting for the functionality of human auditory system, which consists of the peripheral auditory system and the central auditory system. The former is comparatively better understood than the complex central auditory system. For speech COCA analysis, we extract the basic features from digital speech signals leading to a fundamental representation that shares two basic aspects with the human auditory system: A logarithmic dependence on signal power; and a simple bandwidth-to-center frequency scaling so that our frequency resolution is better at lower frequencies. The so-called mel-frequency cepstral coefficients (MFCCs) can loosely represent the human auditory response, except for part of the outer ear, which is critical for sound localization and loudness accuracy. The sound energy is received by the **mechanoreceptors**, and the displacement of the inner hair cells triggers the nerve impulses (Mather, 2006). For detailed description of MFCCs, see (Deller, Hansen, & Proakis, 2000).

To reveal the semantic meaning of an audio signal, analysis over a much longer period is necessary, usually from one second to several tens seconds (Wang, Liu, & Huang, 2000). Feature stacking or vector 'concatenation', as one of the temporal feature integration methods, is by and large a popular means to combine the information from several short time

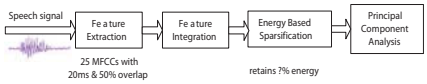

Figure 2: Preprocessing pipeline for COCA of speech. Feature extraction is normally followed by feature integration, so as to obtain features at longer time scales. Energy based sparsification aims at reducing the intrinsic noise and getting sparse representations. PCA projects features onto a base of cognitive processes. A subsequent ICA can identify the actual ray coordinates and source signals.

features (e.g. $20ms$) into a long time feature. This method has been introduced in detail in (Feng & Hansen, 2007). Here the basic MFCCs are 25-dimensional extracted from speech pieces of $20ms$ long with 50% overlap, hence the stacked feature will be $25 * N$-dimensional representing long time scale $20ms * (N+1)/2$.

Sparse representations can be achieved by energy based sparsification (EBS). EBS is a simple way to filter out the weak signals, and it emulates the **detectability** and **sensory magnitude** from perceptual principles (Mather, 2006). For auditory perception only the signals reaching the postsynaptic cell's threshold will lead to the cell firing (Reisberg, 2006). Therefore sparsification is done by thresholding the stacked features, and only coefficients with superior energy are retained, and the rest is set zero.

Principal Component Analysis (PCA) as an orthogonal linear transformation technique, is often used for dimensionality reduction, while the most variance of the data is remained. PCA is known as latent semantic analysis (LSA) in textual information analysis. The semantic content of a document is approximated as the word usage, and is represented as vectors in a semantic space; and the position in the space serves as the semantic indexing (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). Thus it is fully automatic and not syntactic analysis based, but corpus based. The low-dimensional space transformed by LSA from high-dimensional space is regarded as the base for the cognitive processing (Kintsch, 2001). It has been proved that LSA can provide good simulations of human cognitive processes. Here we adopt PCA as the knowledge base of COCA analysis.

Models

In attempt to compare the resulting group structure of unsupervised learning with human cognitive activity reflected by supervised learning of human labels, the unsupervised and supervised learning models should share similarities with respect to the model structure. Furthermore both should allow sparse linear ray-like features. In the previous study, we modified MFA to the unsupervised and supervised ICA-like density models. The independent hypothesis is reflected by the density models in an implicit way. To carry out and emphasize the significance of independency in COCA, we introduce a new pair of models. Since the Bayesian classifier is mis-

classification error rate optimal, here our chosen models are based on Bayesian classifiers: naive Bayes and mixture of Gaussians.

The unsupervised learning model comprises ICA and a naive Bayes classifier. ICA is first applied to the features to recover source signals. Then a naive Bayes classifier, which assumes that the known probabilistic density distribution of the source within each class is Gaussian, will be responsible for revealing the model classification results. To keep the consistency of using Bayesian classifier and Gaussian model, we select mixture of Gaussians (MoG) as the supervised learning model to model the class-conditional probabilities. This is a simple protocol for checking the cognitive consistency: Do we find the same representations when we train them with and without using 'human cognitive labels'?

Unsupervised Learning Model

We introduce ICA into the unsupervised learning model to recover both the mixing coefficients and the original independent sources from the essentially independent sparse features. The vectors defined by the mixing coefficients can be regarded as a set of line-based class indicators in the subspace, to classify samples based on their locations. The typical algorithms for ICA use centering, whitening and dimensionality reduction as three preprocessing steps to reduce the complexity of the algorithm. Since PCA, which achieves these three steps, has already been included in the COCA preprocessing pipeline, we only need to apply ICA directly on the PCA coefficients. Here a noise free ICA model is applied:

$$\mathbf{Y} = \mathbf{AS}, \quad \mathbf{S} = \mathbf{WY}, \tag{1}$$

where $\mathbf{Y}$ is the k-dimensional observation matrix; $\mathbf{A}$ is the mixing matrix with dimension k-by-p; $\mathbf{W}$ is the unmixing matrix; and $\mathbf{S}$ is the matrix of p independent sources, which are assumed non-Gaussian. Without losing generality, we assume the total no. of sources (k) is the same as the dimension of the observation $\mathbf{y}$ (p) in the following experiments, hereby $\mathbf{W} = \mathbf{A}^{-1}$. ICA is able to estimate both the mixing matrix $\mathbf{A}$ and the sources $\mathbf{S}$. This is done by either maximizing the non-Gaussianity of the calculated sources or minimizing the mutual information.

To reveal the performance of unsupervised learning model in classification tasks, we input the recovered source signals with the corresponding manual labels to a naive Bayes classifier, due to the independency of the sources. This is referred to as unsupervised-then-supervised learning scheme.

As the name suggests, the naive Bayes classifier is based on Bayes' theorem:

$$p(\mathbf{C}_i|\mathbf{s}) = \frac{p(\mathbf{s}|\mathbf{C}_i)p(\mathbf{C}_i)}{\sum_i p(\mathbf{s}|\mathbf{C}_i)p(\mathbf{C}_i)}, \tag{2}$$

where $p(\mathbf{C}_i)$ denotes the i^{th} class prior; $p(\mathbf{s}|\mathbf{C}_i)$ is the likelihood of the $\mathbf{C}_i$; and $p(\mathbf{C}_i|\mathbf{s})$ is the posterior of the i^{th} class given data $\mathbf{s}$: $\mathbf{s} = (s_1, \ldots, s_p)^T$. Naive Bayes assumes the independency of input feature for each class, the likelihood in Equation (2) can be simplified as:

$$p(\mathbf{s}|\mathbf{C}_i) = \prod_{n=1}^{p} p(s_n|\mathbf{C}_i), \tag{3}$$

where each $p(s_n|\mathbf{C}_i)$ is modeled as univariate Gaussian distribution $N(\mu_{ni}, \sigma_{ni}^2)$.

For label prediction, we apply the $\mathbf{W}^{train}$ learnt from training set to new data $\mathbf{Y}^{new}$, in order to recover their sources $\mathbf{S}^{new}$. Afterwards, the trained naive Bayes classifier with a set of Gaussian parameters (means and variances) will be used on $\mathbf{S}^{new}$ to predict the labels of new data.

Supervised Learning Model

In this content, the supervised learning model is intended to represent human decisions, therefore we expect it to be a flexible model. The MoG is invoked, as one of the Bayesian classifier family. It follows Bayes' theorem as well. MoG is applied directly to the preprocessed features ($\mathbf{y}$), thus the likelihood is

$$p(\mathbf{y}|\mathbf{C}_i) = \sum_j p(\mathbf{y}|j, \mathbf{C}_i)p(j|\mathbf{C}_i), \tag{4}$$

where $p(\mathbf{y}|j, \mathbf{C}_i) = N(\mathbf{y}|\mu_{ji}, \Sigma_{ji})$, and $p(j|\mathbf{C}_i)$ is the mixing parameters in class $\mathbf{C}_i$. The parameters μ_{ji}, Σ_{ji} are estimated from the training set via the standard Expectation-Maximization algorithm. For simplicity, we assume the covariance matrices to be diagonal. Note that although features are independent within each mixture component due to the diagonal covariance matrix, the mixture model does not factorize over features. The MoG is capable of modeling arbitrary dependency structures among features (Bishop, 1995) if the number of mixture components is sufficiently large. On the other hand, a MoG with many mixture components is prone to overfitting, and will most likely not generalize well. In our experiments, we vary the number of mixture components, and select models according to the classification accuracy. Observations are assigned to the class having the maximum posterior probability. Maximum A *Posteriori* (MAP) criterion aims at maximizing the posterior $p(\mathbf{C}|\mathbf{y})$ rather than the likelihood $p(\mathbf{y}|\mathbf{C})$.

Experiments

The experimental data were gathered from TIMIT database (Garofolo et al., 1993). TIMIT collects reading speech from 630 native American English speakers. Each speaker reads 10 sentences in total, and each sentence lasts approximately 3s. We have several labels for the utterances that we think as cognitive indicators, labels that humans can infer given sufficient amount of data. Here phoneme, gender, age and speaker identity classification are concerned.

Experimental Design

The sentences have been manually labeled with phonetic symbols: 60 phonemes in total; and the age information of

the speakers has also been recorded. We have carefully selected a sufficient amount of data to reach the computational limits of the PC (Intel Pentium IV computer with 3GHz and 2GB of RAM), in the meanwhile we have guaranteed that the data represent the general information of the database. We chose 46 speakers with equal gender partition, and speech signals covered all 60 phonemes, including vowels, fricatives, stops, affricates, nasals, semivowels and glides. To simplify the classification problem, we pre-grouped phonemes into 3 large categories: vowels, fricatives and others. The ages of the TIMIT speakers are not evenly distributed: around 60% speakers are within 21 to 30 years old; and about 30% within age 60 to 72. The ages of the chosen speakers located in the range 21 to 72. Wherefore like phoneme classification, we pre-grouped ages into 4 sets to keep an approximate even population distribution among sets: from age 21 to 25; 26 to 29; 30 to 59; and 60 to 72, both endpoints were included in the set.

The unsupervised and supervised models were compared in a set of experiments: we stacked the basic time scale features into several longer time scales, and sparsified the stacked features with different degrees to test the consistency of the comparison. In the meanwhile of the performance comparison, we also anticipated to find out the role of the time scale. In a particular condition (a certain time scale and sparsification level), the same features have been used in the above mentioned four classification tasks for both unsupervised and supervised learning models, and the difference among four classifications was the class-label information input to the naive Bayes classifier and the MoG.

Following the preprocessing pipeline, we first extracted 25-dimensional MFCCs from speech signals. The 0^{th} order MFCC, which represents the total energy of each short time frame, was also included. To study the role of time scale, we stacked the basic features into a variety of time scales, from basic time scale up to above 1s (20, 100, 150, 300, 500, 700, 900 and 1100ms). The degree of sparsification was controlled by thresholds leading to the retained energy from 100% to 65%. The sparsification was carried out on the normalized stacked MFCCs. PCA was then carried out on stacked and sparsified features, and dimensionality of the features was reduced. For features at longer time scales than 20ms, their dimensions were reduced to 100, and the dimension of the features at the basic time scale remained the same.

The signals from the first 6 sentences of each of the 46 speakers were used as the training set, and were processed following the preprocessing pipeline. The outcomes were input into the unsupervised and supervised models respectively. The ICA algorithm provided us with the unmixing matrix $\mathbf{W}^{train}$, and the sources $\mathbf{S}^{train}$ were consequently recovered in unsupervised learning. Afterwards the sources were input to the naive Bayes classifier together with training set labels to estimate the parameters of the independent univariate Gaussians. For prediction we preprocessed the test set, which consisted of the rest 4 sentences of the 46 speakers, following the same procedure. The $\mathbf{W}^{train}$ was applied to the test set to recover the sources $\mathbf{S}^{test}$. Whereafter the naive Bayes classifier predicted the labels of the test set based on the test sources. We have used the exact same training and test sets for the supervised learning model as for the unsupervised one, so as to exclude the comparison bias introduced by data. MoG model estimated a set of Gaussian distributions from the training set along with the manual labels, and fulfilled the label prediction on the test set. Different number of mixtures was selected based on the classification tasks and the time scales. Both models provided us with a set of predicted labels and a set of posterior probabilities for both data sets.

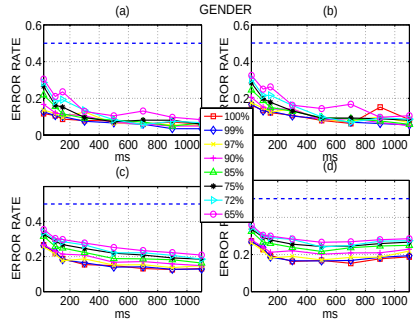

Figure 3: Error rates as a function of time scales for different thresholds in gender classification. (a), (b): Training and test error rates of supervised MoG; (c), (d): Training and test error rates of unsupervised model, respectively; The 8 curves represent feature sparsification with retained energy from 100% to 65%. The dashed lines are the baseline error rates for random guessing. Results indicate that the relevant time scale locates within $300 \sim 500ms$.

Results Comparison

A set of 64 experiments has been carried out in different conditions, i.e. 8 time scales and 8 sparsification levels, for each classification task.

Error Rate Comparison Representations of unsupervised and supervised learning on both training and test sets have been investigated. Here let us first focus on the classification error rates. Figure 3 shows the error rates of gender classification. Plot (a) and (b) are the training and test error rates of MoG separately, whereas (c) and (d) are the training and test error rates of unsupervised learning (ICA+naive Bayes). 8 curves in each panel represent the 8 EBS levels. The tendency of the curves indicates that gender information could be modeled at $300 \sim 500ms$, which coincides with the conclusion of our previous research on gender classification (Feng & Hansen, 2007). The figure also shows that high degree of sparsification, e.g. 65%, degraded the classification accuracy.

Phoneme, age and speaker identity classifications have also

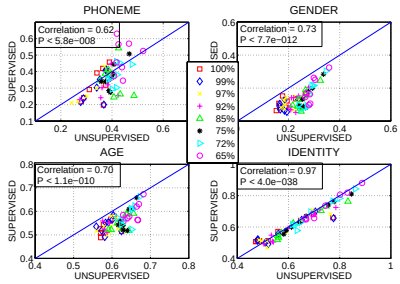

Figure 4: Correlation between test error rates of supervised and unsupervised learning models on four classification tasks: phoneme, gender, age and speaker identity. Solid lines indicate $y = x$. Correlation coefficient and P value for each classification are shown.

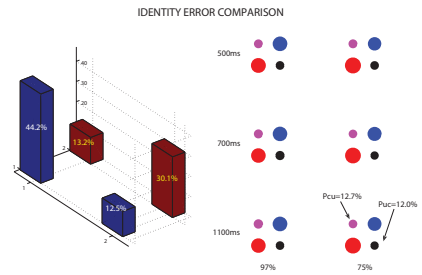

Figure 5: Sample-to-sample test error correlation between supervised and unsupervised learning on identity classification. On the right-hand side, rows represent time scales and columns stand for sparsification degrees. The bottom left circle in each subplot represents P_{cc}, both correctly classified portion by two models; top right shows the both wrongly classified portion P_{uu}. The diagonal circles show the disagreement of two models in making decision: P_{cu} upper left; P_{uc} lower right. On the left-hand side, the histogram summarizes this comparison in all 64 experiments.

been studied, which used the same feature set with different labels indicating the human performance on various cognitive tasks. The results were aligned with those in (Feng & Hansen, 2007) on phoneme and speaker identity classification: first, similarity between supervised and unsupervised learning representations on both tasks was observable; secondly, phonemes were best modeled at short time scale, and speaker identity could be discovered at a longer time scale, such as $> 1s$. Age classification gave similar characteristics on performance comparison, and the recommended time scale lies between gender ($300 \sim 500ms$) and identity ($> 1s$).

To have a close look at the comparison w.r.t. recognition error rates, we measured the correlation of the test error rates. High correlation between the error rates of the two schemes indicated similarity of the representations, shown in Figure 4. The correlations of all tasks were distinguished, while for identity classification: data located nearly along $y = x$, with correlation coefficient $\rho = 0.9660$, and $p < 4.04 \times 10^{-38}$.

Sample-to-Sample Error Comparison In order to reconfirm the finding and to account for the patterns of making decisions for both models, we further computed the error correlation on a sample-to-sample base.

First we computed both correctly classified sample rate by unsupervised and supervised models for the test set of a given task r_{cc}, both wrongly classified sample rate r_{uu}, and the disagreement of two models: correctly classified by supervised model, but wrongly classified by unsupervised model r_{cu}, vise versa i.e. r_{uc}. The total error rates of both models are defined as r_{sup} standing for supervised model; and r_{usup} for unsupervised model. To eliminate the bias caused by total error rate of each model, we thus introduced a new set of rates:

$$R_{cc} = \frac{r_{cc}}{(1-r_{sup})(1-r_{usup})}, \quad R_{uu} = \frac{r_{uu}}{r_{sup}r_{usup}},$$
$$R_{cu} = \frac{r_{cu}}{(1-r_{sup})r_{usup}}, \quad R_{uc} = \frac{r_{uc}}{r_{sup}(1-r_{usup})}. \tag{5}$$

The first row in Equation 5 gives the rates for the matching case; whereas the second row shows the rates of mismatching. Finally to keep the rates as percentages, we normalized them by their summation:

$$P_{ij} = \frac{R_{ij}}{\sum_{mn}(R_{mn})}, \quad m,n = (c,u). \tag{6}$$

Figure 5 shows the degree of matching between the supervised and unsupervised learning models of the test set in speaker identity classification. On the right-hand side, six subplots show the results at a certain time scale and sparsification. In the subplot, the lower left circle refers to the normalized both correctly classified rate by unsupervised and supervised learning: P_{cc}; upper right one stands for P_{uu}. The diagonal circles show the disagreement of two schemes in making decisions: P_{cu} upper left; P_{uc} lower right. The area of each circle represents the portion in percentage, and they sum to 1. The plot reveals that to what degree representations derived from supervised and unsupervised learning match, and how well they match with human labels (the ground truth). On the left-hand side, results of all 64 experiments are summarized into a histogram. In total unsupervised and supervised learning match $44.2 + 30.1 = 74.3\%$, and the matching sits within $P_{cc} + P_{uu} \in [67.9\% \;\; 89.2\%]$ for individual cases. The large percentage allocating on the off-diagonal, indicates high correlation between supervised and unsupervised learning.

Posterior Probability Comparison So far we have seen that the unsupervised and supervised learning models bear close correspondency at the level of error rates and sample-to-sample classification. A more detailed comparison can be obtained by considering the posterior probabilities on the

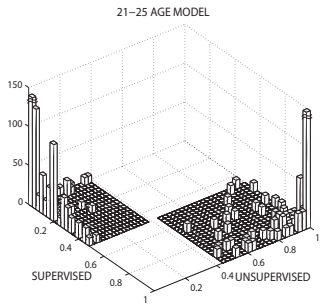

Figure 6: Posterior probability comparison. This figure provides the histograms of the posterior probabilities on the test set, provided by the unsupervised and supervised models for the [21 25] age set in the matching case.

sample-to-sample base. By this means we can measure how the decision certainties match between two models when the final predictions are the same (both correct and both wrong). We chose one experiment from the age classification (700*ms* time scale with 72% remaining energy). Figure 6 presents the posterior probability comparison of unsupervised and supervised models for the 21 to 25 age set. The data shown in the figure belonged to this set. If two models are the exact match, we expect that the posterior probabilities locate along the diagonal of the histograms with high distribution at $(1,1)$ in the coordinate system, which corresponds to the correct decisions by both models, and at $(0,0)$ referring to the wrong decisions by two models. The matching in this case was around 52.5%, with 787 at $(1,1)$ and 769 at $(0,0)$.

Conclusion

With the purpose of understanding the exploitation of statistical regularities in human cognitive activity, we investigated the Cognitive Component Analysis. The protocol we designed to test the cognitive component hypothesis, is to compare the performance of unsupervised learning, which reveals statistical regularities, and supervised learning of manual labels, which loosely represents human cognitive activity. As an extension of our previous work, we employed a new pair of unsupervised and supervised learning models, i.e. ICA followed by naive Bayes and mixture of Gaussians.

With the new models in hand, we have studied the COCA of speech relevant cognitive indicators: phoneme, gender, age and speaker identity. The comparison of the classification performance has been carried out at three levels: error rate level; sample-to-sample level; and the more detailed posterior probability level. The comparisons provided us with the evidence that supervised and unsupervised learning indeed lead to similar representations. Hence it has strengthened our assumption that human cognitive activities are based on statistical regularities, and statistical independence is one of them.

Acknowledgments

This work is supported by the Danish Technical Research Council, through the framework project 'Intelligent Sound' (STVF No. 26-04-0092), www.intelligentsound.org.

References

Barlow, H. (1989). Unsupervised learning. *Neural Computation, 1*, 295–311.

Bell, A. J., & Sejnowski, T. J. (1997). The 'independent components' of natural scenes are edge filters. *Vision Research, 37*, 3327–3338.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. OXFORD University Press.

Deerwester, S. C., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*, 391–407.

Deller, J. R., Hansen, J. H., & Proakis, J. G. (2000). *Discrete time processing of speech signals*. IEEE Press Marketing.

Feng, L., & Hansen, L. K. (2005). On low level cognitive components of speech. In *Proc. international conference on computational intelligence for modelling* (Vol. 2, pp. 852–857).

Feng, L., & Hansen, L. K. (2006). Phonemes as short time cognitive components. In *Proc. icassp* (Vol. 5, pp. 869–872).

Feng, L., & Hansen, L. K. (2007). Cognitive components of speech at different time scales. In *Proc. cogsci* (pp. 983–988).

Field, D. J. (1994). What is the goal of sensory coding? *Neural Computation, 6*, 559–601.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., & Dahlgren, N. L. (1993). The darpa timit acoustic phonetic continuous speech corpus cdrom. In *Nist order number pb91-100354.*

Hansen, L. K., Ahrendt, P., & Larsen, J. (2005). Towards cognitive component analysis. In *Akrr'05*.

Hansen, L. K., & Feng, L. (2006). Cogito componentiter ergo sum. In *Proc. ica* (pp. 446–453).

Hoyer, P., & Hyvrinen, A. (2000). Independent component analysis applied to feature extraction from colour and stereo images. *Network: Comput. Neural Syst., 11*, 191–210.

Kintsch, W. (2001). Predication. *Cognitive Science, 25*, 173–202.

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience, 5*, 356–363.

Mather, G. (2006). *Foundations of perception*. Psychology Press.

Reisberg, D. (2006). *Cognition: Exploring the science of the mind*. W.W.Norton & Company.

Wang, Y., Liu, Z., & Huang, J. (2000). Multimedia content analysis using both audio and visual clues. *IEEE Signal Processing Magazine, 17*.

Appendix H

Cognitive Components of Speech

Submitted for publication in *Artificial Intelligence Journal* 2008. Authors are Ling Feng and Lars Kai Hansen.

Cognitive Components of Speech

Ling Feng and Lars Kai Hansen
Department of Informatics and Mathematical Modelling
Technical University of Denmark

Abstract

COgnitive Component Analysis (COCA) is proposed as a means of investigating the consistency of statistical regularities in a signaling ecology and human cognitive activity. An unsupervised learning algorithm performs cognitive component analysis if it leads to a grouping of data, which is well-aligned with a group structure resulting from human cognitive activity, i.e., the two classifications agree on a majority of scenarios. In our approach we represent human cognitive processes by a *supervised* classification rule based on a set of (human) manually derived labels. This leads to a testable hypothesis and a simple research question: *Do unsupervised component analysis and the supervised proxy for a human lead to comparable classifiers?* The set of computational models we investigate are based on the observation that 'natural' sounds such as speech are represented in as sparse independent components in the brain. Thus our hypothesis is basically ecological: We speculate that features which are essentially independent in a reasonable ensemble, can be efficiently coded using a sparse independent component representation. In addition we are interested in speech at different time scales searching for possible hidden 'cognitive structure'. The basic features are short-time mel-frequency weighted cepstral coefficients, assumed to model the basic frequency based representation of the human auditory system. A simple temporal feature integration method, namely feature stacking is applied to obtain features at longer time scales. Simple energy based filtering is used to achieve a sparse representation. We design a series of experiments to reveal the statistical regularities and their relation to human cognitive processes. By measuring the correlation of classifier output at three levels: error rates level, sample-to-sample prediction level and posterior probability level, we show that the resulting representations of unsupervised (ecological) and supervised (human) are indeed very well aligned, hence lending additional support to the cognitive component hypothesis.

This work is supported by the Danish Technical Research Council, through the framework project 'Intelligent Sound' (STVF No. 26-04-0092), www.intelligentsound.org.

The Cognitive Component Hypothesis

We are interested in computational models of human cognition. In particular models of the data analytic processing pipelines invoked by the human brain – 'the cognitive architecture'. Human cognition is the result of an extensive evolutionary optimization process of the data analytic pipeline under biological, computational, and statistical constraints. Humans make decisions in real time, hence, the evaluationary 'objective function' has traded off precision and speed. The primary function of human cognition is to integrate percepts with memory to prepare for and execute, action. The resulting human cognitive system can model complex multi-agent scenery, and uses a broad spectrum of cues for analyzing perceptual input and for identification of individual signal processes.

It is believed that an evolutionary optimized brain is capable of exploiting statistical regularities while making inference on appropriate actions (Barlow, 1989). *Statistical independence* is likely to be such regularity. Barlow posed the question: *What is the source of the extensive and well-organized knowledge of the environment implied by the possession of an cognitive map or working model?*, and provided a partial answer by noting that an unsupervised learning algorithm can provide us with a factorial code of independent visual features; and that our visual feature detectors reduce redundancy in a representation based on variables which are approximately statistically independent. Related ideas have been investigated, e.g., in (Pearlmutter & Hinton, 1986).

The exploration for independent components in relevant natural ensemble statistics, has been carried out for more than a decade. Bell and Sejnowski extracted 'independent components' from an ensemble of natural scenes, and also demonstrated the detection of edges in natural images by linear filters derived by independence assumptions. Furthermore they anticipated the predictive power of abstract unsupervised learning techniques (Bell & Sejnowski, 1997). Additional studies of independence in primary sensory systems include (Hoyer & Hyvrinen, 2000) on visual feature extraction from images, and (Lewicki, 2002) on natural sound coding, where sounds were categorized into three distinct classes: non-harmonic environmental sounds; harmonic animal sounds; and speech having both harmonic vowels and non-harmonic consonants. The factorial code derived from independence potentially involves of all orders, however most studies only focus on second order statistics to de-correlate outputs of a set of feature detectors. One of the most dramatic demonstrations is the work by Eizaburo Doi et al. on independent components in color imagery (Doi, Inui, Lee, Wachtler, & Sejnowski, 2003). In natural color images form and color features are broadly independent, hence, segregate in the minimum redundancy representations. Similar results have been obtained for video data (Hateren & Ruderman, 1998). The segregation of representations of form, color and motion in the brain are well documented.

Sparseness, like independence, has the property of reducing computational complexity. While independence lead to simple factorial codes, sparseness simply reduces the number of computations by eliminating weak evidence. In sensory coding, 'sparse distributed' coding was invoked and proved to be near optimal in representing natural scenes in the visual system (Field, 1994). These studies have shown that the sensory information is encoded in an energy efficient manner by a small number of active neurons at a given point of time. Field has argued for the importance of sparseness in which the above mentioned statistical independent feature detector is activated as rarely as possible, in line with Barlow's

'Minimum Entropy coding' principle. The principle of sparse coding has a history of more than three decades. It has been suggested and studied from different viewpoints and reasons. A current review on the sparse coding (Olshausen & Field, 2004) has summarized its advantages.

Theoretical and experimental studies of the sensory input encoding in cerebral cortex have been carried out in great detail. Especially, in the visual system, see e.g. (Field, 1994; Olshausen & Field, 1996; Bell & Sejnowski, 1997) receptive field properties seem to match sparse representations found in computational models. In the auditory system, the receptive field properties of auditory nerve cells invoke a strategy of sparse independent manner to represent natural sounds, see e.g. (Lewicki, 2002; Olshausen & O'Connor, 2002).

Combining research on perceptual representations and engineering models of blind signal separation, the field of Independent Component Analysis (ICA) emerged in the mid90's (Comon, 1994). Although generalizations have been proposed for non-linear mixing, most current ICA models are based on linearly mixed source signals. The structure of a linear mixture depends on the statistical properties of the mixed source signals. Let us broadly characterize the histogram of a signal as sparse, normal, or dense. Sparse signals are distributed so that small and very large signals are more prominent compared to signals produced by the normal distribution, while a dense signal has less small and less large signals. These three types of signals lead to vastly different mixtures as seen in Fig. 1. The upper panel shows a typical appearance of a sparse source mixture. The sparse signals are made of a few samples with relatively very large magnitude in a background of a mass number of small or weak signals. When mixing such independent sparse signals in a simple linear manner, we will most likely end up with a 'ray structure', which we consider emblematic for our COCA approach. ICA algorithms have been invented based on various forms of higher order statistics (beyond second order). The linear mixture can be recovered (mixing vectors and source signals) if maximally one of the source signals is normal. ICA is by now a commonly used statistical tool for analysis of, e.g., audio, video, medical and text data (Hyvrinen, Karhunen, & Oja, 2001).

The Hypothesis

Our hypothesis is ecological: We assume that features that are essentially independent in a context defined ensemble can be efficiently coded using a **sparse independent** component representation. COgnitive Component Analysis (COCA) is defined as the process of unsupervised grouping of generic data such that the ensuing group structure is well-aligned with that resulting from human cognitive activity, see (Hansen, Ahrendt, & Larsen, 2005).

We represent the results of human cognitive activity in an intuitive way. Since the mechanisms of human cognitive activity are still not fully understood, to quantify cognition may seem ambiguous and may also be considered way too ambitious. However human behavior, as the direct consequence of cognition, contains rich phenomenology, and is easier to access and model than human cognition. As to speech relevant cognitive activities, we focus attention on auditory human behavior in the context of speech perception, speech context understanding, and judgment based on speech. Hence we explain the broad term human cognition in our specific studies by a classification rule, i.e. based on a set of manually obtained labels we train a classifier using supervised learning. The manually obtained labels reflect human judgment of a given task. *The question is then reduced to*

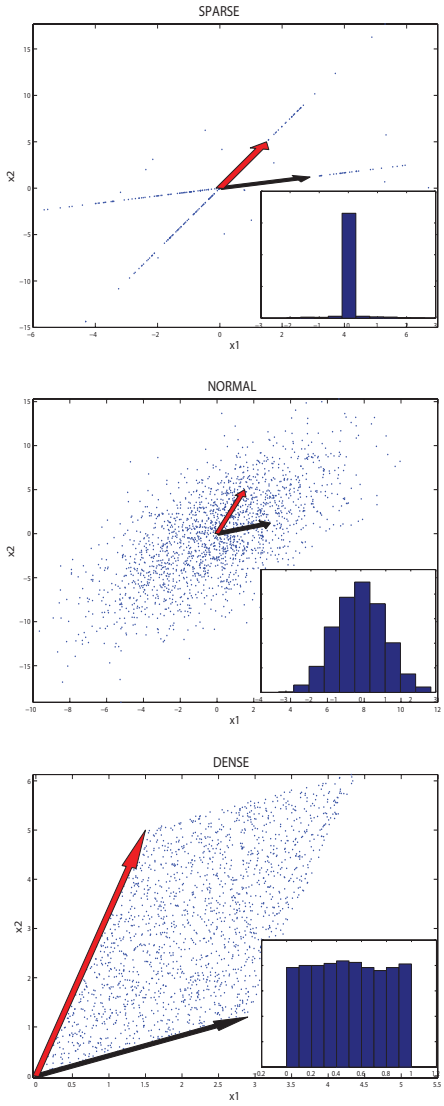

Figure 1. Prototypical feature distributions produced by a two-dimensional linear mixture based on two sources with sparse (upper panel), normal (middle), or dense histograms (lower panel), respectively. A sparse signal contains relatively few large magnitude samples on a background of weak signals (see inlet in the upper panel), hence, produces a characteristic ray structure in which rays are defined by the vector of linear mixing coefficients: One ray for each sparse source.

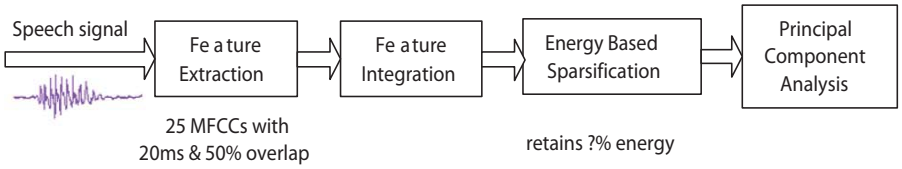

Figure 2. Preprocessing pipeline for COCA of speech. Feature extraction is normally followed by feature integration, so as to obtain features at longer time scales. Energy based sparsification aims at reducing the intrinsic noise and getting sparse representations. PCA projects features onto a base of cognitive processes. A subsequent ICA can identify the actual ray coordinates and source signals.

looking for similarities between the representations in supervised learning (of human labels) and unsupervised learning that simply explores the statistical properties of the domain. The supervised model only captures the specific aspect of human cognition. If the representations provided by both unsupervised learning and supervised learning methods are well aligned with each other, we take it as the evidence for our hypothesis.

Data Preparation – Preprocessing Pipeline

Directly working on raw data is usually not optimal and practical, and the representative information for various tasks needs to be extracted in an efficient way. This step is normally called feature extraction. To emulate how brain processes speech signals in the early stages, we design the preprocessing pipeline, starting from feature extraction, shown in Fig. 2. The design of the preprocessing procedures focuses on two aspects: on one hand, the flow-chart is built in respect to standard techniques used in speech signal processing; on the other hand, the choice of techniques are in connection with the human auditory system response.

Feature Extraction

To choose features which are capable of representing the information that human ear perceives, we need to briefly delve into the physiology of human ear, psychophysics and psychoacoustics.

Feature extraction is influencing to the overall performance. Features are basically extracted from short time scales. The frame size may depend on applications, in other words features at different time scales may contain different information. A small frame size may result in a noisy estimation; on the contrary a long frame size may lose the appropriate information in need. Later we will discuss the rule of the time scale. To represent speech signals for machine speech analysis, spectral features of fairly low dimensionality are usually used. These $20 - 30$ dimensions of features are usually uncorrelated. The basic features in COCA analysis are extracted from a digital speech signal leading to a fundamental representation that shares similarities with the human auditory system.

From the physiology of human ear viewpoint, a nerve fibre will response to a broad range of frequencies, however the maximum response happens only if the sound frequency

matches this nerve fibre's *characteristic frequency*. The phase and intensity of a sound wave are claimed to be reflected by the pattern of firing in auditory nerve fibres. Therefore some researchers claim that loosely speaking, the ear is a Fourier analyzer, where each sound can be decomposed to a collection of sine frequency components (Mather, 2006). The psychophysical studies of frequency masking declare that humans use a function like a band-pass filter to perceive signals, select frequencies within the bandwidth, and remove the rest. Studies following this line support a view that the function of the human auditory system for signal perception is fulfilled by a bank of bandpass filters from low frequencies (e.g. 20 Hz) to high frequencies (e.g. 16 kHz). Based on these theoretical standpoints, so-called mel-frequency cepstral coefficients (MFCCs) are invoked. MFCCs are designed as perceptually weighted cepstral coefficients, since the mel-frequency warping emulates human non-linear frequency perception of sound. MFCCs have been developed for speech processing, e.g. speech recognition and speaker recognition (Reynolds & Rose, 1995). However MFCCs recently have been popular in many other areas, such as music genre classification (Ahrendt, Meng, & Larsen, 2004), audio similarity measure (Arenas-Garca et al., 2007) and instrument classification (Nielsen, Sigurdsson, Hansen, & Arenas-Garca, 2007), etc.

Due to the non-stationarity of speech, basic features need to be extracted from audio signals in, e.g. $10 - 40$ *msec*, in which period the signal is assumed stationary. Here we name these short-time features the basic features in COCA analysis. The computation of MFCCs is based on the short-time time-frequency analysis. MFCCs decompose signals into broad spectral channels, and compress the loudness of the signals. The block diagram for computing MFCCs is given in Fig. 3. The fast Fourier transform (FFT) transforms the convolution relationship between excitation sequence and the vocal system impulse response into production in the frequency domain; and the logarithm, afterwards, provides us with the addition of these two. The mel-frequency warping changes the frequency scale from linear to mel-scale, which attempts to mimicry the non-linear human pitch perception. The mel-frequency warping is realized by a bank of bandpass filters, termed critical band filters. A few types of filters can be used, such as triangular shaped filters, hanning filters and hamming filters. Here triangular shaped filters are in use, and the center frequencies spacing of filters follows mel-scale. Loosely speaking, the critical band filters represent the frequency resolution of the peripheral human auditory system, and they also reflect the auditory system in a way that signals passing through different critical bands are processed independently (Glasberg & Moore, 1990). Finally discrete cosine transform (DCT) brings us to the mel-cepstrum. For detailed description, see (Deller, Hansen, & Proakis, 2000). All in all, MFCCs share two aspects with the human auditory system: A logarithmic dependence on signal power and a simple bandwidth-to-center frequency scaling so that the frequency resolution is better at lower frequencies. Therefore they can loosely represent the human auditory response, except for part of the outer ear, which is critical for sound localization and loudness accuracy.

Feature Stacking

Feature integration is referred to the process of constructing features at a longer time scale than the basic ones, so as to obtain discriminative information for a given task.

Among all the feature integration methods, a simplest approach is to stack short-time features into a long vector. Feature stacking has been used in audio retrieval and indexing

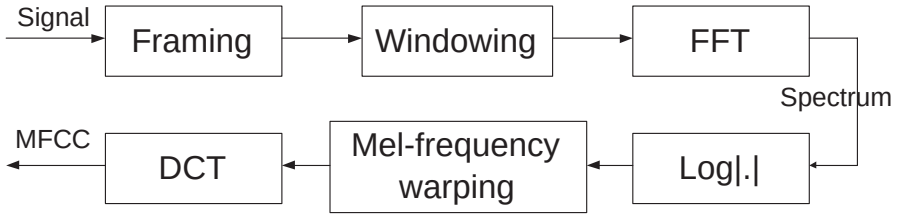

Figure 3. Block diagram of mel-frequency cepstral coefficients

to obtain long-term spectral characteristics of short-time MFCCs (Slaney, 2002).

Fig. 4 illustrates the stacking procedure used in COCA analysis.

1. Truncate speech signal into short time frames, $20ms$ which corresponds to 320 samples at 16 kHz sampling frequency. A certain overlap between two adjacent frames is set, e.g. 50%;

2. Since the side lobes of the rectangular window spectrum cause signal power to 'leak' into other frequencies, a hamming window is applied to each frame;

3. d-dimensional MFCC is extracted from each frame, e.g. a 25-dimensional feature vector;

4. According to the time scale, the MFCCs from the first N frames are stacked into one $d * N$-dimensional vector;

5. Repeat 4 with the next N short time frames (without overlap) until all the short time frames are stacked (and exclude the residual).

The $d * N$-dimensional features extracted with 50% overlap among short-time frames, represent speech information at $20msec*(N+1)/2$ long time scale. In the later stage of the preprocessing pipeline, dimensionality reduction algorithm will be used to select the most important dimensions.

Energy Based Sparsification

The receptive field properties of auditory nerve cells invoke a strategy of sparse independent manner to represent natural sounds. Hence we here carry out the energy based sparsification (EBS), and it is also meant to emulate the cognitive process: 'attention', in a way that strong (loud) signals win awareness.

Attention is the ability to concentrate on one event in the surrounding while ignoring the other events. Concentrating on one person speaking or one conversation in a noisy environment is one example. The cocktail party problem involves both attention concentration and attention shift, e.g., you shift your attention while somebody outside your conversation calls your name. Here we 'borrow' the concept and interpret 'attention' simply as focusing on signals with large magnitude (energy) in the background of many weak signals.

EBS is a simple way to filter out weak signals, and it emulates the **detectability** and **sensory magnitude** from perceptual principles (Mather, 2006). Detectability in percep-

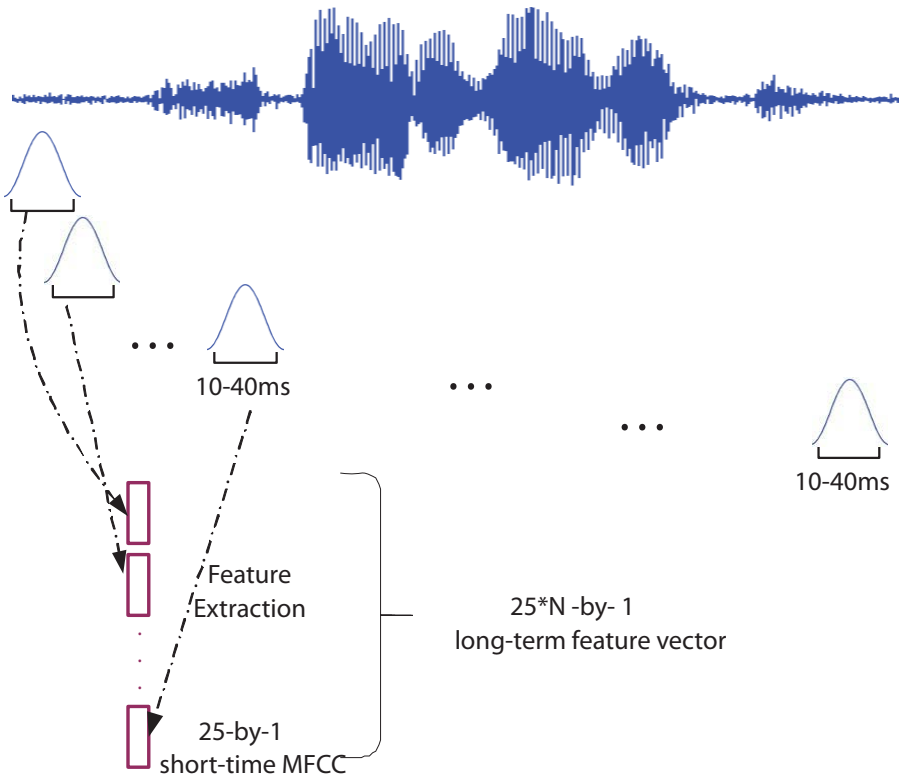

Figure 4. Speech feature extraction and stacking.

tual principles means the ability of sensory organs to detect the environmental stimulus. It depends highly on the intensity of the stimulus and the variability of neural signals as well. When a stimulus is sensed, the *neurotransmitters* between dendrites and terminal endings of neurons are generated. If the neurotransmitters are larger than a certain threshold of dendrites, the cell will fire, and in turn send signal to other neurons (Reisberg, 2006). In short, the relationship between the intensity of a stimulus and the dendrites threshold will influence the detectability. Therefore sparsification is done by thresholding the stacked features. Since MFCC coefficients are energy based, the thresholding is applied directly on the amplitude of the coefficients, and only coefficients with superior energy than the threshold are retained, and the rest is set zero. In this study, the thresholds are set empirically.

Principal Component Analysis

COCA is built based on a generalization of principal component analysis (PCA) based 'latent semantic analysis', originally developed for information retrieval on text (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). PCA is an orthogonal linear transformation technique. It is often used for dimensionality reduction, which transforms and projects the data to a new coordinate system with lower dimensions, and in the meanwhile remains the most variance of data.

In textual information analysis, latent semantic indexing (LSI) or latent semantic analysis (LSA) assumes that semantic content of the text, can be reflected by the sum of the meaning of words it includes. This assumption successfully avoids the complex syntactic problems, and converts the semantic indexing to a corpus based problem. For information retrieval, the task is to match the words of queries with words of documents or the conceptual content of documents. Since the words in a search query are not always included in the aiming documents, or in some cases the words of query may be covered in some irrelevant documents, where different meanings of the words have been refereed, to discover the latent semantics is indispensable. Normally, text data are formed as a large term-document matrix. The terms are the representative words in the documents. The matrix will be decomposed by some statistical machine learning techniques, and also will be projected into a low-dimensional 'semantic' space from the original high-dimensional space. In this low-dimensional space, words are seen as points, and meaning is represented as vectors (Salton, 1989; Deerwester et al., 1990). Therefore the position in the space is served as indexing, and documents having similar or common topics locate close to each other in the space. The resulting low-dimensional space is regarded as the basis for all cognitive processing (Kintsch, 2001). Some cognitive scientists believe that the performance of LSA resembles humans performance in the way meaning is represented. Since LSA has human-like performance in text analysis, we envision that it can as well be used to get the relevant basis for cognitive related tasks, e.g. speech perception. It has been proved that in some cases, LSA can provide good simulations of human cognitive processes alone, and in other cases it is often operated as base for cognitive processes. Here we adopt this well-understood concept, PCA/LSA, as the knowledge basis of COCA analysis, and use other ways to transform it.

To grasp the essential information and discard the redundancy (e.g. noise) in the data, singular value decomposition (SVD) is invoked to select the most informative and important dimensions in a sense that maximal amount of variance is retained. The mathematical express of SVD on data matrix X is

$$\mathbf{X} = \mathbf{U}\Lambda\mathbf{V}^T, \tag{1}$$

where $\mathbf{X}$ is a m-by-n matrix; $\mathbf{U}$ is a m-by-m orthonormal matrix; Λ is a m-by-n matrix with singular values along the diagonal; and $\mathbf{V}$ is a n-by-n orthonormal matrix. The dimensionality of data is reduced by projecting the data to the first k principal components ($k < m$):

$$\mathbf{Y} = \mathbf{U}_k^T\mathbf{X} = \Lambda_k\mathbf{V}^T. \tag{2}$$

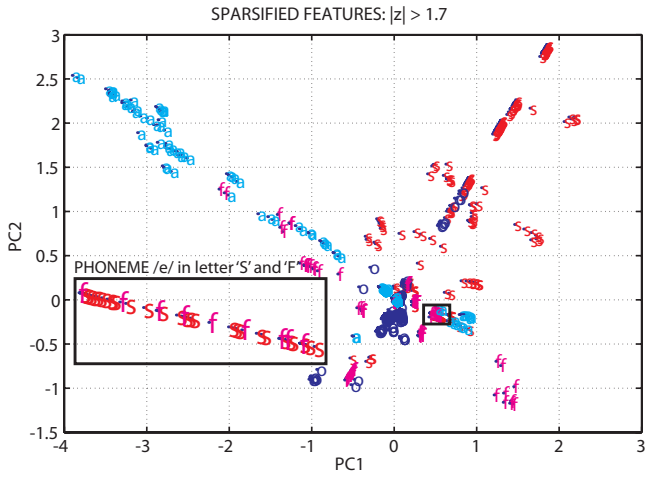

Figure 5. The latent space is formed by the first two principal components of data consisting of four separate utterances (letters): 's', 'o', 'f', 'a'. The structure clearly shows the sparse component mixture, with 'rays' emanating from the origin (0,0). The ray embraced in a rectangle contains a mixture of 's' and 'f' features, a cognitive component associated with the vowel /e/ sound.

Low-level Cognitive Components

From our previous work on speech signals (Feng & Hansen, 2006), we have reported the preliminary findings of ICA ray structure related to phonemes in a relatively small data set. The unsupervised learning method, ICA has transformed the orthogonal basis vectors derived from PCA based LSA on phoneme features. Fig. 5 illustrates the phoneme relevant ray structure at the basic time scale. This analysis was carried out on four letters: 's', 'o', 'f' and 'a' from TIMIT database, which is a reading speech corpus designed for automatic speech recognition systems. Cognitive components of /e/ phoneme opening 's' and 'f' are identified. We speculate that these phoneme-relevant cognitive components contribute towards the well-known basic 'invariant cue' characteristics of speech (Blumstein & Stevens, 1979). The theory of acoustic invariants points out that perceived signals are derived as stable phonetic features despite of the different acoustic properties produced by different speakers. Moreover Damper has shown that although the speech signal may vary due to coarticulation, the relation between key features follows a consistent and invariant form (Damper, 1998).

Here we will focus on revealing 'invariant cue' in a simple situation: one particular phoneme in the subspace of different phonemes pronounced by one person. The low-level COCA has been carried out on two letters: 'g' and 't', and their phonetic symbols are: /dgi:/ and /ti:/. Theoretically they share a same phoneme: vowel /i:/. According to the 'invariant cue' theory, the phonetic features derived from different words are invariant to their environments (here we mean the surrounding phonetic features, e.g. /dg/ and /t/) and different trials. We used 'g' and 't' letters from TIMIT letter database. The

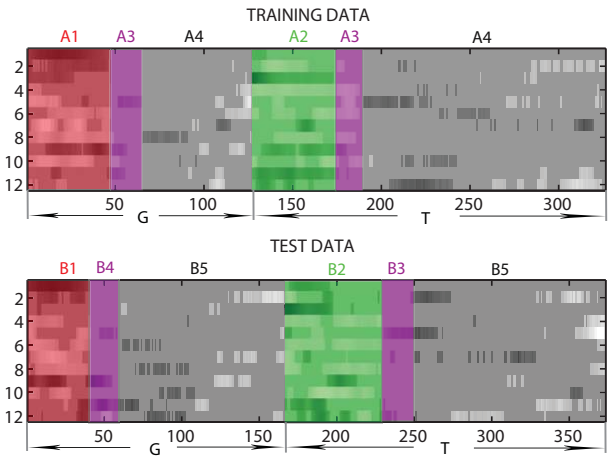

Figure 6. Sparsified mel-frequency cepstral coefficients. The sparsification threshold is $z = 1.2$, which keeps 68% energy. Both training and test sets are given. Several regions are marked corresponding to different phonetic groups. Their locations in scatter plots are shown in Fig. 7.

first trials of these two letters were used as training set, while the second as test set. 12-dimensional MFCCs were extracted from 333 samples at 10 kHz with 95% overlap between adjacent short time frames. These MFCCs used hamming windows in time domain and triangular mel-filters in the absolute log frequency domain. MFCCs from two letters were concatenated. Threshold for sparsification was set to keep 68% of total energy in the remaining coefficients. Fig. 6 shows the sparsified MFCCs, note that samples obtaining zero energy have been removed. Afterwards, PCA found the eigenvectors, hence features were projected into principal components. Fig. 7 gives the scatter plot of the training and test samples in the subspace of two principal components. The 'ray-structure' is striking. We have studied the samples in the small data set, and found out their corresponding locating areas in the temporal development of MFCCs. The plots are divided into several regions, marked from A1 to A4 for training set, and from B1 to B5 for test set. Since the scatter plot is in 2 dimensions, the regions indicating different phonetic features could have overlap. Their area coverage has been roughly marked in the time domain shown in Fig. 6. Loosely speaking, region A1 indicates the phonetic features related to /dg/; A2 corresponds to /t/; A3 are the transient parts from both /dg/ to /i:/ and /t/ to /i:/; while region A4 is the 'invariant cue' we are looking for: the vowel /i:/ existing in both letters 'g' and 't'. Similar to training set, B1 is the group of MFCC samples from phoneme /dg/; B2 refers to /t/; B3 is the transient components from /t/ to /i:/; while B4 is the transient from /dg/ to /i:/; finally B5 represents the common /i:/ sound from both 'g' and 't'. The results confirm to those included in (Feng & Hansen, 2006), and also the LSA analysis on text data, which shows that text having similar semantic meanings locate near one another in the semantic space. This conclusion can be translated, in the sense of phonemes, to such that samples

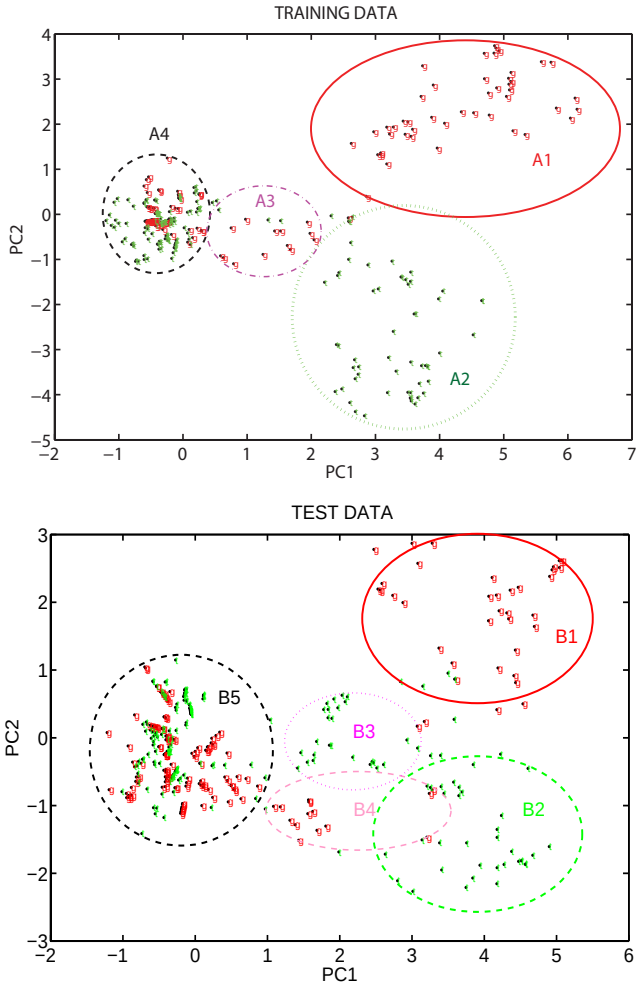

Figure 7. Scatter plot of training and test set in the subspace of the 1st and 2nd principal components. Samples are labeled as g and t, indicating their affiliations. The 'ray-structure' is observable. By studying temporal locations of these samples, we allocate them as regions in sparsified mel-frequency cepstral coefficients (Fig. 6).

sharing similar phonetic characteristics locate close to each other in the phonetic space.

However invariant cues are hard to identify, and investigators on this subject have different views: some believe that phoneme is the fundamental unit in speech perception, and the invariant characteristics are derived from these units; some believe that invariant

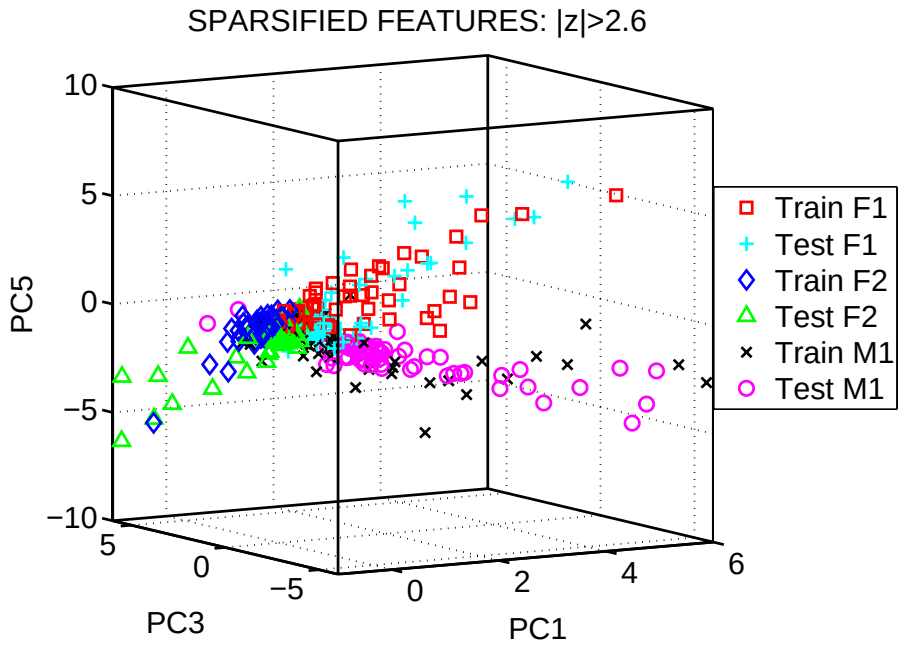

Figure 8. We focus on text-independent speech recognition. This experiment was carried out on two female and one male speakers, denoted as: F1, F2 and M1. All data points from both training and test data sets are shown in the space of the 1st, 3nd and 5th principal components. There is an evident ray structure corresponding to a generative ICA model based on linear mixing of sparse sources. We see that the ray structure is solely determined by the speaker identity. Rays from the training and test sets are closely aligned.

cues are not static but dynamic, and therefore can not be associated with a single phoneme; some even question about the phoneme being the fundamental unit, and claim that the unit in speech perception is not unique, but depends on the focus of attention of the brain. One of the reasons is that they believe speech perception is based on syllables or words, and hence phonemes are not perceived, but perhaps inferred from the perceived syllables or words (Dusan & Rabiner, 2005). Nevertheless, the 'invariant cue' has been revealed here

based on the first view that invariant property is derived from phonemes.

Experiments involving labels related to speaker identification also provided the signature of linear 'ray'-structures, shown in Fig. 8 (Feng & Hansen, 2005). This experiment was carried out on a subset of our in-house speaker recognition database, ELSDSR. The figure gives the signature structure of COCA analysis on speaker recognition task (each subject is enrolled with a different set of reading material). The basic features were 12-dimensional MFCCs, and they were integrated into longer time scale, say $1sec$. Sparsification helped us get significant sparse ICA 'ray' structure emanating from origin of the coordinate system. Similar experiment has also been done based on the same text among speakers. The results showed us both phoneme-like characteristics of the 'ray' structure, and speaker dependent property. We stipulate that this effect is an interaction between the text content and the speaker identity.

In both phoneme and speaker identity recognition which are cognitive relevant tasks from speech signals, we end up with linear structures. Is linearity related to perceptually distinguishable categories? The discussion on linear correlations in speech signal and the locus equation is still on-going (Sussman, Fruchter, Hillbert, & Sirosh, 1998).

As shown, during the itinerary of exploring spoken cognitive components, we have already reported the generalizable phoneme relevant components at a time scale of $20 \sim 40ms$, and the generalizable speaker specific components at an intermediate time scale of $1000ms$. *In this paper we will further expand on our findings in speech by applying COCA on speech features at various time scales. We will systematically investigate the performance of unsupervised and supervised learning, and test whether the tasks are learned in equivalent representations. The positive answer will hence indicate the consistency of statistical regularities and human cognitive processes.*

Models

Unsupervised learning typically regards input signals as a set of random variables, and it aims at investigating and extracting regularities in the input vectors. It is sometimes called self-organization. On the contrary, supervised learning usually deals with training patterns which consist of pairs of input signal and the desired output. In our experiments the desired outputs are discrete manual labels, indicating the supervised learning is for classification tasks. The supervised learning learns from the training data to establish a function, and later uses the function to predict outputs. Moreover the function should be generalizable so that it is also able to predict unseen situations. The mechanism of supervised learning is consistent with concept learning in human psychology.

Here we will examine the performance of unsupervised COCA on speech signals, and systematically compare it with performance of supervised learning method. To keep the comparison consistent, all the experiments will follow the basic COCA preprocessing scheme shown in Fig. 2, therefore the unsupervised and supervised models will work on the same feature representations for a variety of tasks in speech understanding. The similarity measure will be carried out at the level of error rates comparison, at the sample-to-sample error correlation level, and at the posterior probability level as a proxy for the similarity. Later we will encounter the models: the unsupervised learning model with unsupervised-then-supervised scheme, followed by the supervised learning model. Finally the experiment design will be given together with comparison results.

Having the comparison of the unsupervised and supervised learning in mind, we need to have two models which share similarities w.r.t the model structure. Moreover both models should allow sparse linear ray-like features. The Bayesian classifier which assumes a known probabilistic density distribution for each class, has been widely used and is misclassification error rate optimal. Besides, Bayesian theory conveys the *likelihood principle* in perception, and it infers optimally under conditions of uncertainty. Thus it is capable of revealing plausible perceptual decisions (Feldman, 2004). Here we use two Bayesian classifiers: Naive Bayes and Mixture of Gaussian (MoG).

For the unsupervised learning model we first apply unsupervised ICA only on the features. After recovering source signals, we add the label information to a naive Bayes classifier, which assumes the distribution of the source within each class is Gaussian. To keep the consistency of using Bayesian classifier and Gaussian model, we choose MoG as the supervised learning model. This is a simple protocol for checking the cognitive consistency: Do we find the same representations when we train them with and without using 'human cognitive labels'?

Unsupervised Learning

As introduced in the previous section, our two hypotheses of cognitive component analysis are sparsity and independency: we presume that the human auditory system follows the sparse coding rule; and if the sparse features are essentially independent, we can use ICA to recover both mixing coefficients and original independent sources.

As we know typical algorithms for ICA use centering, whitening and dimensionality reduction as preprocessing steps in order to reduce the complexity of the algorithm. PCA is normally used to achieve these steps. Algorithms for ICA include infomax, FastICA and JADE, but there are many others also. Since in the preprocessing pipeline we have applied PCA on stacked and sparsified MFCC features, we directly apply ICA algorithm on PCA coefficients without dimensionality reduction.

The component y_i of the observation vector $\mathbf{y} = (y_1, \ldots, y_k)^T$ are generated by summing independent sources $\mathbf{s} = (s_1, \ldots, s_p)^T$ with different mixing weights $a_{i,n}$:

$$y_i = a_{i,1}s_1 + \ldots + a_{i,j}s_j + \ldots + a_{i,p}s_p. \tag{3}$$

The generative formula of the noise free ICA model is

$$\mathbf{Y} = \mathbf{AS}, \tag{4}$$

where $\mathbf{Y}$ is the k-dimensional observations; $\mathbf{A}$ is the mixing matrix with dimension k-by-p; $\mathbf{S}$ is the matrix of p independent sources which are assumed non-Gaussian. ICA aims at estimating both the mixing matrix $\mathbf{A}$ and sources $\mathbf{S}$. This is done by either maximizing the non-Gaussianity of the calculated sources or minimizing the mutual information.

The original sources can be recovered by

$$\mathbf{S} = \mathbf{WY}, \tag{5}$$

where we assume the total number of sources (k) is the same as the dimension of the observation $\mathbf{y}$ (p) in the following experiments, hereby $\mathbf{W} = \mathbf{A}^{-1}$ is the unmixing matrix, and the $\mathbf{A}$ and $\mathbf{W}$ matrices are therefore square.

To reveal the performance of unsupervised learning in classification tasks, we first train the unsupervised model using only the features (principal components) $\mathbf{y}$ to recover the sources $\mathbf{S}$. Since sources are independent, then naive Bayes classifier can be applied to sources with training set labels. This is also referred to as unsupervised-then-supervised learning scheme.

As the name suggests, the naive Bayes classifier is based on Bayes' theorem:

$$p(\mathbf{C}_i|\mathbf{s}) = \frac{p(\mathbf{s}|\mathbf{C}_i)p(\mathbf{C}_i)}{\sum_i p(\mathbf{s}|\mathbf{C}_i)p(\mathbf{C}_i)} \tag{6}$$

where $p(\mathbf{C}_i)$ denotes the i^{th} class prior; $p(\mathbf{s}|\mathbf{C}_i)$ is the likelihood of $\mathbf{C}_i$; and $p(\mathbf{C}_i|\mathbf{s})$ is the posterior of the i^{th} class given data $\mathbf{s}$: $\mathbf{s} = (s_1, \ldots, s_p)^T$.

The naive Bayes assumes that data variables are independent, therefore the likelihood in equation (6) can be simplified as:

$$p(\mathbf{s}|\mathbf{C}_i) = \prod_{j=1}^{p} p(s_j|\mathbf{C}_i), \tag{7}$$

where each $p(s_j|\mathbf{C}_i)$ is modeled as univariate Gaussian distribution $\mathcal{N}(\mu_{ji}, \sigma_{ji}^2)$.

For the classification problem, we apply the $\mathbf{W}$ learnt from training set to new data $\mathbf{Y}^{new}$, and recover its sources $\mathbf{S}^{new}$. Afterwards, the trained naive Bayes classifier with a set of Gaussian parameters (means and variances) will be used on $\mathbf{S}^{new}$ to predict labels of the new data.

Supervised Learning

For the supervised learning model, we have chosen a very flexible model, the so-called Mixture of Gaussians (MoG) to capture statistical properties of the cognitive tasks we consider.

$$p(\mathbf{C}_i|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{C}_i)p(\mathbf{C}_i)}{\sum_i p(\mathbf{y}|\mathbf{C}_i)p(\mathbf{C}_i)}, \tag{8}$$

and the likelihood will be modeled by the MoG, i.e.:

$$p(\mathbf{y}|\mathbf{C}_i) = \sum_j p(\mathbf{y}|j, \mathbf{C}_i)p(j|\mathbf{C}_i), \tag{9}$$

where $p(\mathbf{y}|j, \mathbf{C}_i) = \mathcal{N}(\mathbf{y}|\mu_{ji}, \Sigma_{ji})$ denotes Gaussian distribution with mean μ_{ji} and covariance Σ_{ji}, and $p(j|\mathbf{C}_i)$ is the mixing parameters in class $\mathbf{C}_i$. The parameters μ, Σ are estimated from training data of each class via the standard Expectation-Maximization (EM) algorithm. For simplicity, we assume the covariance matrices to be diagonal. Thus the axes of the resulting Gaussian 'blobs' are parallel to the axes of the input data space. Although features are independent within each mixture component due to the diagonal covariance matrix, the mixture model does not factorize over features. The MoG is capable of modeling arbitrary dependency structures among features (Bishop, 1995) if the number of mixture components is sufficiently large. On the other hand, a MoG with many mixture components is prone to overfitting and will most likely not generalize well. In our experiments, we vary the number of mixture components, and select models according to classification accuracy.

The classification first computes the likelihood estimate for each entrained MoG model, and the incoming observation is assigned to the class having the maximum *posterior* probability, which also takes the subjective *priori* into account.

Experiment Design and Results

The experiments were carried out on speech signals gathered from TIMIT database (Garofolo et al., 1993). For each utterance we have several labels that we think as cognitive indicators, labels that humans can infer given sufficient amount of data. While each sentence lasts approximately $3s$, there are totally 10 sentences reading by each speaker, we will investigate the performance of features at time scales ranging from the basic $20ms$ to about $1s$. The cognitive labels we focusing on here, are phoneme, gender, age and speaker identity. The total speech covers 60 phonemes, and the age of speakers ranges from 21 to 72. We have carefully selected a sufficient amount of data to reach the computational limits of the PC (Intel Pentium IV computer with $3GHz$ and $2GB$ of RAM), and in the meanwhile we have guaranteed that the chosen data represent as much of the breadth of the available information in TIMIT as possible w.r.t. the information of interest. we chose 46 speakers with equal gender partition, and the age of these speakers covers the whole age range, and speech signals cover all 60 phonemes.

Following the preprocessing pipeline, we first extracted 25-dimensional MFCCs from original speech signals with hamming windows in the time domain and triangular filters in the mel-frequency domain. Within these 25 dimensions, the so-called 0^{th} order MFCC is also included, which represents the log-energy of each short time frame. To investigate various time scales, we stacked the basic features into a variety of time scales, from $20ms$ scale up to $1100ms$ (20, 100, 150, 300, 500, 700, 900 and $1100ms$). Energy based sparsification was used afterwards as a means to reduce the intrinsic noise and to obtain sparse signals. The degree of sparsification was controlled by a threshold in the experiments, and by changing the threshold leading to a retained energy from 100% to 65%, we examed the role of sparsification. PCA was then carried out on stacked and sparsified features, and dimensionality of features was reduced. For features having longer time scales than 20 ms, their dimensions were reduced to 100, and the dimension of features at the basic time scale remained the same, i.e. 25.

We divided 10 sentences from each speaker into two sets: the first 6 sentences were enrolled into the training set, the rest was put into the test set. In the training phase, after preprocessing of features, the training data were input into unsupervised and supervised models respectively. The ICA algorithm estimated the unmixing matrix $\mathbf{W}^{train}$, and source signals $\mathbf{S}^{train}$, which were used in the naive Bayes classifier together with training set labels to estimate parameters of the independent univariate Gaussians. For prediction, we preprocessed the test set following the same procedure. The $\mathbf{W}^{train}$ was applied to the test set to recover the sources $\mathbf{S}^{test}$. Whereafter naive Bayes classifier predicted labels of the test set based on test sources. We have used the exact same training and test set for the supervised model in order to exclude the comparison bias caused by data. Mixture of Gaussian models estimated a set of Gaussian distributions from the training set for each class, and fulfilled the label prediction on the test set by looking at the maximum posterior probabilities of each unknown sample. Both models provided us a set of labels and a set of posterior label probabilities for both data sets.

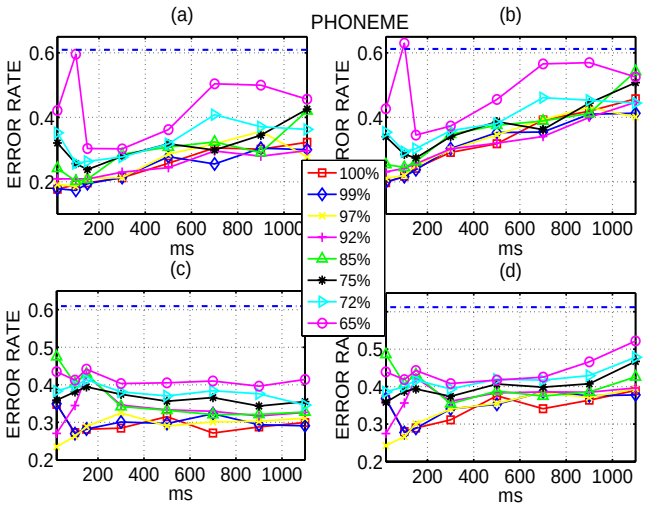

Figure 9. Error rates as a function of time scales for different thresholds in phoneme recognition. (a), (b): Training error rates and test error rates of MoG respectively; (c), (d): Training error rates and test error rates of ICA+naive Bayes. The 8 curves represent feature sparsification with retained energy from 100% to 65%. Dashed lines are baseline error rates for random guessing. Results indicate that phonemes are best modeled at short time scale: around $20ms$.

Phoneme Recognition

We first examined phoneme recognition within the chosen speakers. The 60 phonemes, which are covered by the selected signals, include vowels, fricatives, stops, affricates, nasals, semivowels and glides. To simplify the classification problem, we pre-grouped these phonemes into 3 large categories: *vowels*, *fricatives* and *others*. A set of experiments were carried out in 64 (8 times 8) conditions, i.e. 8 time scales and 8 sparsification levels. We trained with appropriate manual labels in supervised learning models to represent the human observer, and with the unsupervised - then - supervised scheme to represent the 'ecological' grouping. Fig. 9 presents the results of both supervised and unsupervised learning on this task. The two plots (a) and (b) show the training and test error rates of the mixture of Gaussian models separately, whereas (c) and (d) are the training and test error rates of the unsupervised learning (ICA+naive Bayes). These curves in each panel represent the 8 EBS levels. It is obvious that features at longer time scales degraded the performance, which coincides with the conclusion from our previous research that phonemes are best modeled at short time scales (Feng & Hansen, 2006, 2007, 2008). Furthermore sparsification does play a role: with too few energy left, e.g. 65%, the recognition error rates went much higher.

Error Rates Comparison From the above experiments we noticed that the unsupervised and supervised models performed similarly w.r.t recognition error rates. To exam

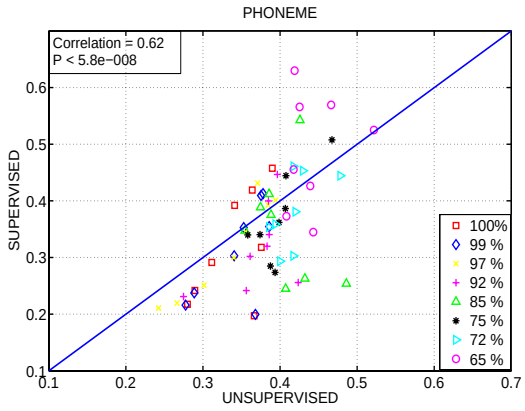

Figure 10. Correlation between test error rates of supervised and unsupervised learning on the phoneme label set. The solid line indicates line $y = x$ in the given coordinate system. The correlation coefficient and P value are given.

how well their representations are correlated, we followed the approach outlined above. In both unsupervised and supervised learning cases we can measure the test performance of the resulting classifiers. High correlation between error rates of the two schemes indicated the similarity of the representations, shown in Fig. 10. The correlation is observable in phoneme recognition task: for the given time scales and thresholds, data scatter in the coordination system, and a large portion of data sit close to line $y = x$, and the correlation coefficient is $\rho = 0.62$, $p < 5.8 \times 10^{-8}$.

Sample-to-Sample Error Correlation In order to reconfirm the finding and to account for the patterns of making mistakes for both models, we further looked into the error correlation on a sample-to-sample base. First we computed both correctly classified sample rate by unsupervised and supervised models for the test set of a given task r_{cc}, the both wrongly classified sample rate r_{uu}, and the disagreement of two models: correctly classified by supervised model, but wrongly classified by unsupervised model r_{cu}, vise versa, i.e. r_{uc}. To eliminate the effect caused by total error rate of each model, we calculated the misclassified rates by each model: r_{sup} and r_{usup}. Whereafter we have the following four rates:

$$\begin{aligned} R_{cc} &= \frac{r_{cc}}{(1-r_{sup})(1-r_{usup})}, \quad R_{uu} = \frac{r_{uu}}{r_{sup}r_{usup}}, \\ R_{cu} &= \frac{r_{cu}}{(1-r_{sup})r_{usup}}, \quad R_{uc} = \frac{r_{uc}}{r_{sup}(1-r_{usup})}. \end{aligned} \tag{10}$$

The first row in Equation 10 gives the rates for the matching case; whereas the second row shows the rates of mismatching. Finally to keep the rates as percentages, they are all normalized:

$$P_{ij} = \frac{R_{ij}}{\sum_{mn}(R_{mn})}, \qquad m, n = (c, u). \tag{11}$$

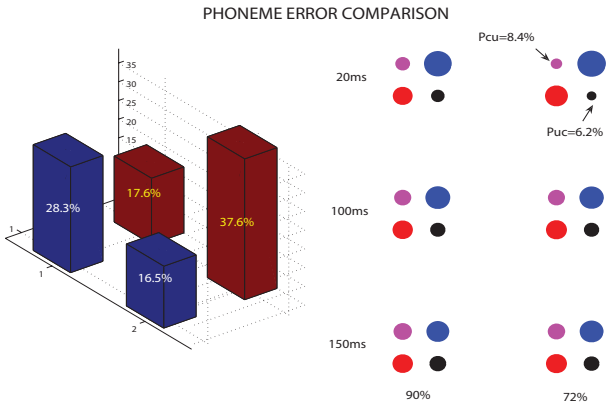

Figure 11. Sample-to-sample test error correlation between supervised and unsupervised learning in phoneme recognition. On the right-hand side, rows represent time scales from $20 \sim 150$ ms and columns stand for different sparsification degrees, corresponding to the retained energy 90% and 72%. The bottom left circle in each subplot represents P_{cc}, both correctly classified portion by two models; top right shows the both wrongly classified portion P_{uu}. The diagonal circles show the disagreement of two models in making decision: P_{cu} upper left; P_{uc} lower right. On the left-hand side, the histogram summarizes this comparison in all 64 experiments.

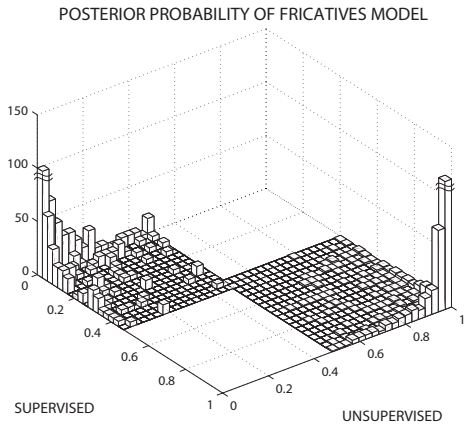

Figure 12. Posterior probability comparison. It shows the histograms of posterior probabilities provided by unsupervised and supervised *fricatives* models on the test set in the matching case. Two highest distributions locate at $(1, 1)$ and $(0, 0)$, which are 678.7 and 440.3 respectively.

Fig. 11 shows the degree of matching (mismatching) between the supervised and unsupervised learning models on the test set. On the right-hand side, six subplots show the results at a certain time scale and sparsification. In the subplot, the lower left circle refers to the normalized both correctly classified rate by unsupervised and supervised learning: P_{cc}; upper right one stands for P_{uu}. The diagonal circles show the disagreement of two schemes in making decisions: P_{cu} upper left; P_{uc} lower right. The area of each circle represents the portion in percentage, and the four areas sum to 1. It presents to what degree representations derived from supervised and unsupervised learning match, and how well they match with human labels (the ground truth). A large percentage allocates on the off-diagonal circles, indicating high correlation between supervised and unsupervised learning. On the left-hand side of this figure, results of all 64 experiments are summarized into a histogram. The locations of bars correspond to the circles in subplots. In total, the matching of both learning schemes in decision making held $37.6 + 28.3 = 65.9\%$. For individual cases, the matching lied in the range of $P_{cc} + P_{uu} \in [35.1\% \;\; 85.4\%]$. The highest correlation happened at $20ms$ time scale and with 72% remaining energy, which was 85.4% consistency between two models; and the corresponding error rates for supervised learning was 35.4%, and 38.6% for unsupervised learning.

Posterior Probability Comparison So far we have seen that there is a close correspondence at the level of error rates and sample-to-sample classification. A more detailed comparison can be obtained by considering posterior probabilities on a sample base. For each sample, we will get a number of posterior probabilities for unsupervised learning, and the same number of posterior probabilities for our supervised learning. We are aiming at comparing the corresponding probabilities provided by both unsupervised and supervised learning methods when they make the same predictions, either they are both right or wrong. This comparison measures the degree of certainty of two models making the same decision. We chose one experiment of the phoneme recognition ($100ms$ time scale with 97% remaining energy) among the 64 experiments mentioned above. Phoneme recognition is within 3 classes, and for label prediction on a new sample, we have three posterior probabilities provided by *vowels*, *fricatives* and *others* models individually for each type of learning. Fig. 12 presents the posterior probability comparison of the *fricative* models when they make the same predictions. If two models are the exact match, we should expect that posterior probabilities locate along the diagonal of the histograms with high distribution at $(1, 1)$ and $(0, 0)$. The matching in this case was around 49.7%.

Gender Detection

Similar experiments have also been carried out in gender detection within the chosen 46 speakers, within which half are female speakers. We used the same experiment setup with the same feature sets, and the only difference lies in the label information. Fig. 13 presents the results of two learning methods on this task. Same as before, (a) and (b) give the training and test error rates of the supervised learning separately, and (c),(d) are the training and test error rates of unsupervised learning. First, we note that error rates were decreasing while time scale was increasing, and somewhere around 500 ms curves tended to converge. Moreover sparsification shows its role again: when a high percentage of features was retained from sparsification, e.g. 100% and 99%, error rates did not change much with the time scale increment, meaning the longer time scales do not assist to enhance the

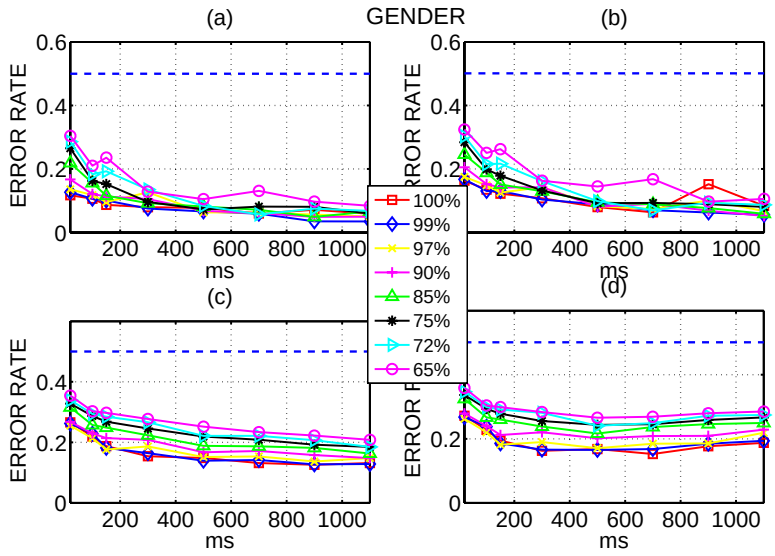

Figure 13. Error rates as a function of time scales for different thresholds in gender detection. (a), (b): Training error rates and test error rates of supervised MoG respectively; (c), (d): Training error rates and test error rates of unsupervised ICA+naive Bayes. The 8 curves represent feature sparsification with retained energy from 100% to 65%. Dashed lines are baseline error rates for random guessing. Results indicate that the relevant time scale is about $300 \sim 500ms$ for this cognitive relevant task.

performance; while with too few energy left, recognition error rates went higher.

Error Rates Comparison High correlation between the error rates of the two schemes indicates the similarity of the representations, shown in Fig. 14. Compared to phoneme recognition, the correlation here is more distinguished in gender detection task: the data tend to follow two trends, for data having light sparsification, they tend to follow a line with a small slope, whereas data having high sparsification degree tend to locate along a line with a deeper slope than $y = x$, and the overall correlation coefficient $\rho = 0.73$, $p < 7.7 \times 10^{-12}$.

Sample-to-Sample Error Correlation Now we have a look at the comparison between unsupervised and supervised learning on the sample-to-sample base. The same as before we computed P_{cc}, P_{uu}, P_{cu} and P_{uc}. Fig. 15 shows the degree of matching (mismatching) between two type of learning on the test set. Again we observed that a large percent of prediction allocates on the off-diagonal circles and bars, which indicates high correlation, and the matching of two models in total was $51.4 + 21.7 = 73.1\%$, and for individual cases, the matching sat in the range of $P_{cc} + P_{uu} \in [61.2\% \;\; 89.9\%]$.

The highest correlation happened at $700ms$ time scale and with 65% remaining energy,

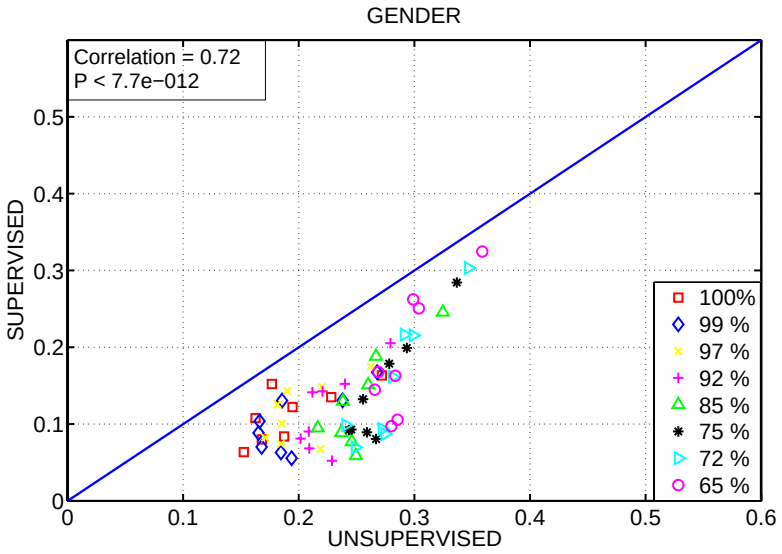

Figure 14. Correlation between test error rates of supervised and unsupervised learning in gender detection. The solid line is $y = x$. The correlation coefficient and P value are given.

which was 89.9% consistency between two models; and the corresponding error rates for supervised learning was 16.8%, and 26.9% for unsupervised learning. From Fig. 13 we learnt that the longer the time scale, the better the classification performance. Nevertheless we do not aim at finding the best performance, rather we should focus on the correlation between two models within the area of recommended time scale for gender detection.

Posterior Probability Comparison To measure the degree of certainty of unsupervised and supervised models making the same decisions, we chose one experiment: $500ms$ time scale with 97% remaining energy. Gender classification is a binary problem, therefore we obtain two posterior probabilities provided by *female* model and *male* model for each type of learning. Fig. 16 presents the posterior probability comparison of the *female* models. The matching was around 74.8% here.

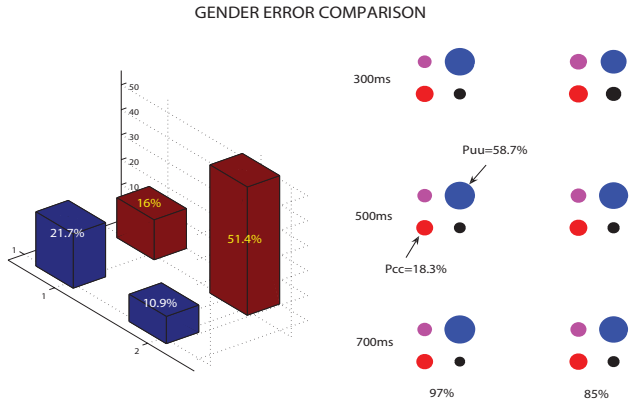

Figure 15. Sample-to-sample test error correlation between supervised and unsupervised learning in gender recognition. On the right-hand side, rows represent time scales from $300 \sim 700$ ms and columns stand for different sparsification degrees, corresponding to the retained energy 97% and 85%. The bottom left circle in each subplot represents both correctly classified portion P_{cc}; top right shows the both wrongly classified portion P_{uu}. The diagonal circles show the disagreement of two models in making decision: P_{cu} upper left; P_{uc} lower right. On the left-hand side, the histogram summarizes this comparison in all 64 experiments.

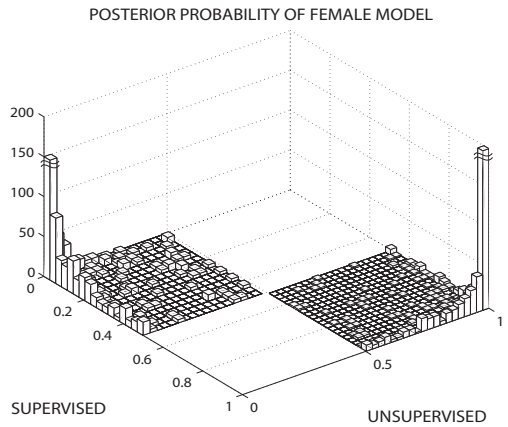

Figure 16. Posterior probability comparison. Figure shows the histograms of posterior probabilities provided by unsupervised and supervised *female* models on the test set in the matching case. Two highest distributions locate at $(1, 1)$ and $(0, 0)$, which are 1153 and 379.2 respectively.

Identity Recognition

The speaker-specific cognitive components have been found out in the previous work, here we will compare the findings from unsupervised learning with representations derived from supervised learning. Similar experiments have been performed on this cognitive tasks within 46 speakers. The corresponding label set were given together with the same feature sets used in the previous experiments. Similar to Fig. 9 and 13, Fig. 17 provides the results of the supervised learning model: mixture of Gaussian models and the unsupervised learning model: ICA + naive Bayes. The error rates were decreasing with the increment of time scales, and the high degree of sparsification caused the increment of error rates. Due to the lack of data, which was around $30s$ per speaker from TIMIT database, longer time scale integration is not feasible. Thus we speculate that speaker identity should be modeled at longer time scale, e.g. $> 1s$.

Error Rates Comparison By regarding the corresponding error rates from unsupervised learning and supervised learning as a point in the coordinate system, we measured the error rate correlation. Fig. 18 reveals the high correlation of these two learning schemes, indicating the similarity of the representations. The correlation is astonishing: all data points locate tightly along $y = x$, which gives a correlation coefficient $\rho = 0.97$, and a P value of $p < 4.0 \times 10^{-38}$.

Sample-to-Sample Error Correlation Extremely High correlation has been shown in the error rate correlation for speaker recognition, now let us study the similarity at a even detailed level, namely on a sample-to-sample base. Following the same procedure, we first calculated the four rates P_{cc}, P_{uu}, P_{cu} and P_{uc}, and presented them in a few subplots, and summarized the overall results in a histogram. Fig. 19 gives the degree of matching (mismatching) between two learning schemes on the test set. We once again observed that the off-diagonal circles and bars overtook the diagonal components, and in total unsupervised and supervised learning matched in $44.2 + 30.1 = 74.3\%$, and the matching sat in the range of $P_{cc} + P_{uu} \in [67.9\% \;\; 89.2\%]$ for individual cases.

Posterior Probability Comparison By investigating posterior probabilities given by a number of models, we measure the degree of certainty/uncertainty. The chosen data set includes 46 speakers, hence it is a 46-class classification problem. For each learning method we have 46 models, which will produce 46 posterior probabilities to make a single prediction. In the test set each speaker had about $12s$ recording. To obtain a reasonable classification result, the feature integration needs to cross speech of $1s$. Hence we had a few test data points for each speaker. Here we chose one experiment: $900ms$ time scale with 92% remaining energy, and show poster probabilities from a number of speaker models. We chose 12 models for posterior probability comparison, and show them in Fig. 20. The plots in three rows represent the best, the moderate and the worse cases. Each subplot gives posterior probabilities of both correct decisions (red *), and both wrong decisions (blue triangle). The perfect match between unsupervised and supervised learning leads to a 'y=x' line. Here we define the matching percentage as the percentage of samples having the difference between unsupervised and supervised posterior probabilities within ± 0.1, and the percentage was ranging from 33.3 % to 85.7 % for the pair-to-pair comparison of all 46 models.

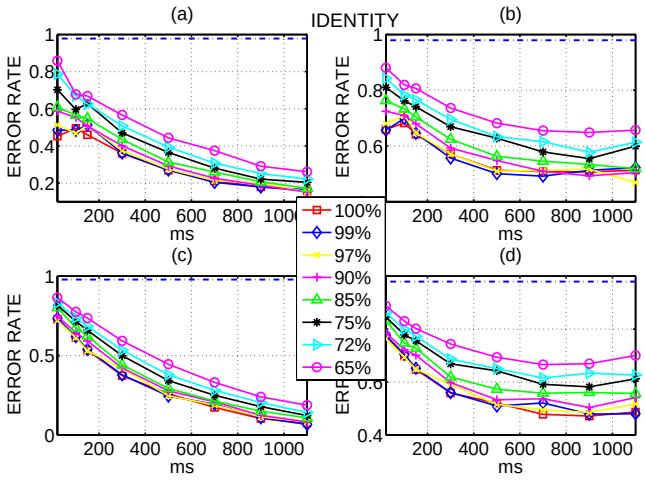

Figure 17. Error rates as a function of time scales for different thresholds in speaker identity recognition. (a), (b): Training error rates and test error rates of supervised MoG respectively; (c), (d): Training error rates and test error rates of unsupervised ICA+naive Bayes. The 8 curves represent feature sparsification with retained energy from 100% to 65%. Dashed lines are baseline error rates for random guessing.

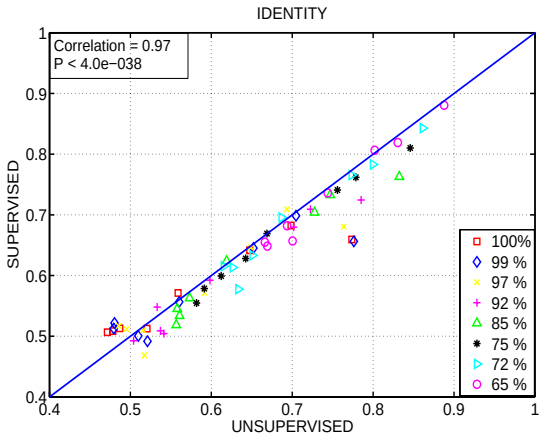

Figure 18. Correlation between test error rates of supervised and unsupervised learning in identity recognition. The solid line indicates $y = x$.

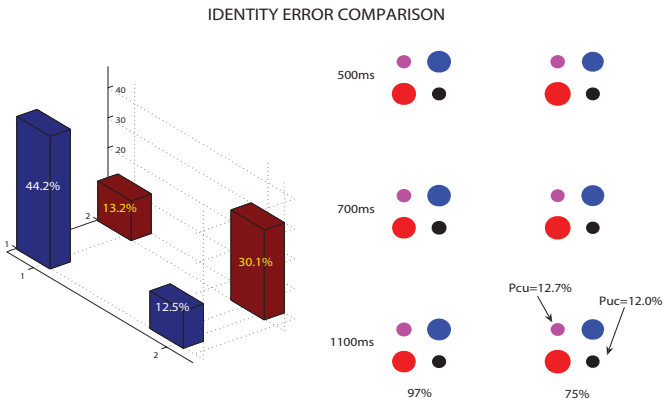

Figure 19. Sample-to-sample test error correlation between supervised and unsupervised learning on speaker ID recognition. On the right-hand side, rows represent time scales from 500 $\sim$ 1100 ms and columns stand for different sparsification degrees, corresponding to the retained energy 97% and 75%. The bottom left circle in each subplot represents P_{cc}, and the top right one shows P_{uu}. The diagonal circles show the disagreement of two models in making decisions: P_{cu} upper left; P_{uc} lower right. On the left-hand side, the histogram summarizes this comparison in all 64 experiments.

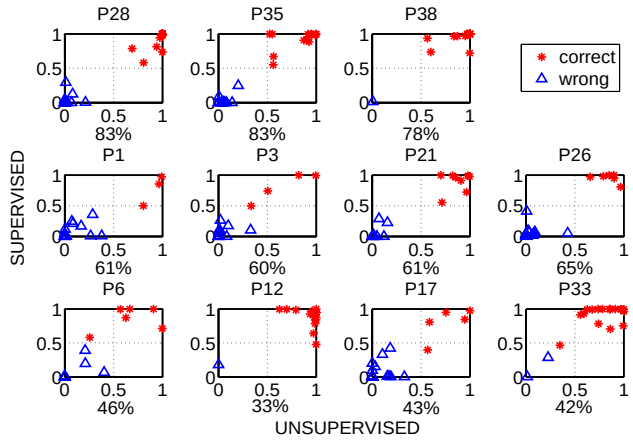

Figure 20. Posterior probability comparison. 12 models are selected, e.g. P28. P1 to P23 are female speakers; the rest is male. Each sub-figure is an unsupervised vs. supervised posterior probability plot on the test set in the matching case. The percentage of matching is given.

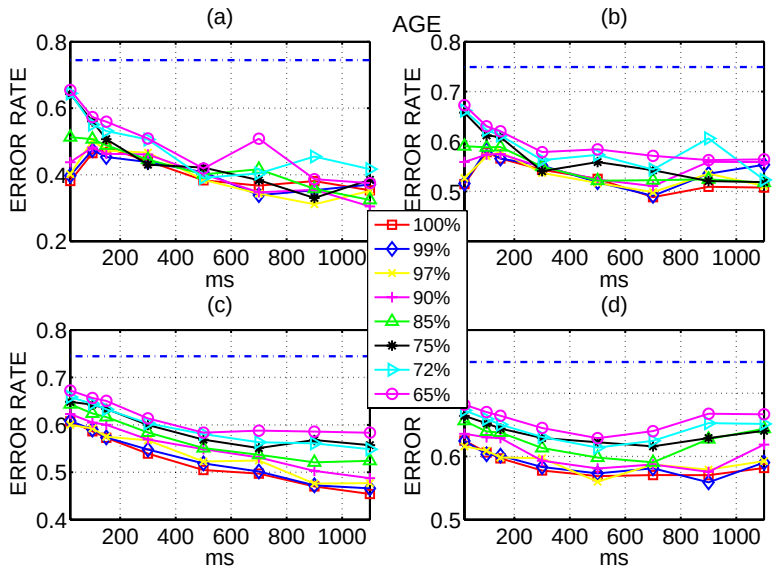

Figure 21. Error rates as a function of time scales for different thresholds in speaker's age detection. (a), (b): Training error rates and test error rates of supervised MoG respectively; (c), (d): Training error rates and test error rates of unsupervised ICA+naive Bayes. The 8 curves represent feature sparsification with retained energy from 100% to 65%. Dashed lines are baseline error rates for random guessing.

Age Detection

Finally we focus our attention on one potential cognitive indicator: age. The age of the TIMIT speakers are not evenly distributed: around 60% speakers are within 21 to 30 years old; and about 30% within age 60 to 72. The age of the chosen speakers located in the range from 21 to 72. Wherefore like phoneme recognition, we pre-grouped ages into 4 sets to keep an approximate even population distribution among sets: from age 21 to 25; 26 to 29; 30 to 59; and 60 to 72, both endpoints were included in the set.

We carried out similar experiments in age detection. Fig. 21 shows the error rates of supervised and unsupervised learning: (a),(b): the training and test error rates of the MoG models; (c),(d): the training and test error rates of ICA+naive Bayes. Similar as identity recognition, when the time scale increased, the error rates decreased. It seems to be saturated around 1s. We speculate that age might be modeled at time scale around 1s.

Error Rates Comparison Following the routine, we tested the consistency of both learning methods at the error rate level. In Fig. 22 we take the error rates from unsupervised learning as the x-value, and error rates from supervised learning as y-value for a certain coordinate system. Data points sat along a line parallel to $y = x$, revealing that supervised

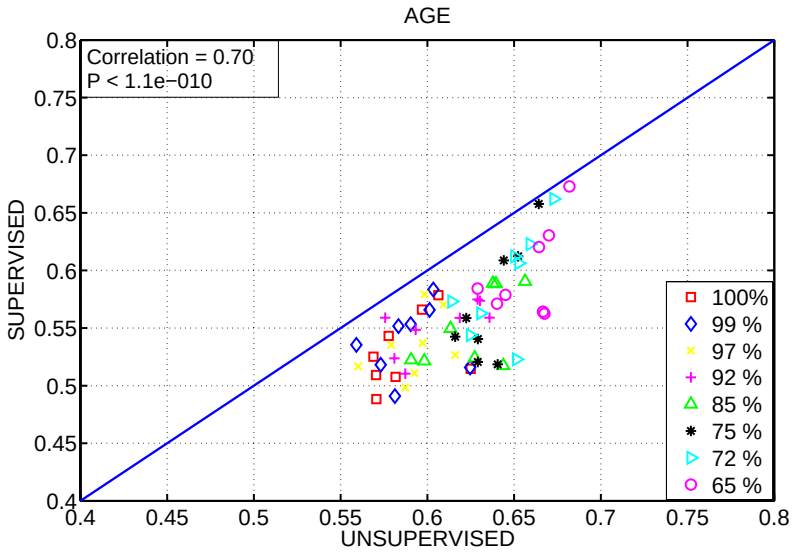

Figure 22. Correlation between test error rates of supervised and unsupervised learning on age detection. The solid line indicates $y = x$. The correlation coefficient is shown together with the P value.

learning always outperforms unsupervised learning in this task, however they are still highly correlated, represented by the correlation coefficient $\rho = 0.7$, and a P value of $p < 1.1 \times 10^{-10}$.

Sample-to-Sample Error Correlation We show the degree of matching between two learning methods performing on the test set at a sample-to-sample prediction basis in Fig. 23. The off-diagonal circles in the subplots are the rates: P_{cc} and P_{uu} in a particular case. The histogram adds the four rates (P_{cc}, P_{uu}, P_{cu} and P_{uc}) in all the difference conditions together. For age detection, the off-diagonal still took over, and in total (the histogram) unsupervised and supervised learning had $32.0 + 28.4 = 60.4\%$ matching, and the matching of a certain case sat in the range of $P_{cc} + P_{uu} \in [42.0\% \;\; 81.1\%]$.

Posterior Probability Comparison A posterior probability tells us how certain of a model in making a particular decision. Since we pre-grouped age information into 4 groups, it became a 4-class classification problem. Here we chose a experiment with $700ms$ time scale and 90% remaining energy. Fig. 24 shows the posterior probability comparison of the young people: *21-25 age* models. The matching in this case was around 44.7%.

The posterior probability comparison of all four tasks lead to a similar conclusion about the certainty matching of two models having the same prediction: when one model was certain (having the posterior probability close to 1 or 0), the other was also certain by having the value of the same order; when they were both uncertain, the degree of uncertainty showed less consistence.

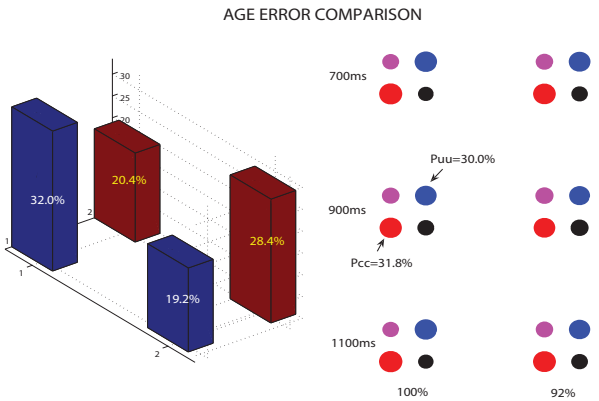

Figure 23. Sample-to-sample test error correlation between supervised and unsupervised learning in speaker's age detection. On the right-hand side, rows represent time scales from $700 \sim 1100\ ms$ and columns stand for different sparsification degrees, corresponding to the retained energy 100% and 92%. The bottom left circle and the top right circle in each subplot represent the matching case: P_{cc} and P_{uu} respectively. The diagonal circles show the mismatching of two models in decision making: P_{cu} upper left; and P_{uc} lower right. On the left-hand side, the histogram summarizes this comparison in all 64 experiments.

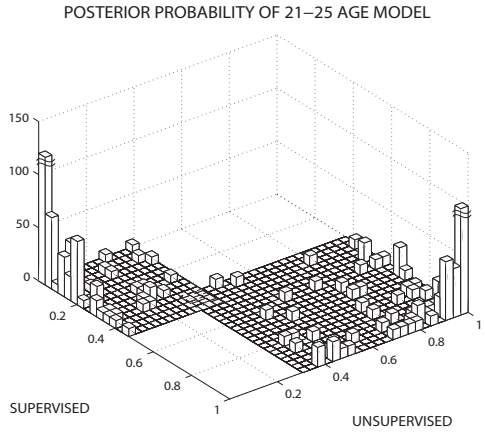

Figure 24. Posterior probability comparison. It shows the histograms of posterior probabilities provided by unsupervised and supervised *21-25 age* models on the test set in the matching case. Two highest distributions locating at $(1, 1)$ and $(0, 0)$ are 404.9 and 520.4 respectively.

Conclusion

With the purpose of examine the consistency of statistical regularities and human cognitive activity, we proposed the cognitive component analysis which has been defined as unsupervised grouping of data so that the resulting group structure is well-aligned with that resulting from human cognitive activity. We have devised a protocol to test the cognitive component hypothesis, i.e. to compare the performance of unsupervised learning, which aims at discovering statistical regularities, and supervised learning, which loosely represents human cognitive activity, under closely matched conditions, so that the only difference is that 'cognitive labels' are used for supervised learning while not for unsupervised learning.

We preprocessed speech following a pipeline starting from feature extraction. The basic features were short time (e.g. $20ms$) mel-frequency cepstral coefficients. MFCC is so far the most well-known and representative feature for human auditory perception, and its design has taken two basic aspects with the human auditory system into account. Feature stacking was used to aggregate features at multiple time scales. Energy based filtering on stacked features led to a sparse distributed representation, which in the meanwhile also reduced the intrinsic noise of speech signals. The variation of intensity effects of speech was reflected in MFCCs by means of magnitude. By thresholding out the coefficients with lower magnitudes, we excluded the portion caused by the physical stimulus containing low energy. SVD based PCA on the preprocessed feature set brought us to the cognitive knowledge base for COCA analysis.

We first extended previous research findings on low-level COCA. Unsupervised learning helped to reveal 'invariant cue' on phoneme data, which is the invariant language units existing in difference environments and different trials. We devised a pair of models to represent the unsupervised and supervised learning. To carry out the statistical independent hypothesis, we employed ICA as the unsupervised learning model followed by naive Bayes to reveal the classification capability of the unsupervised learning model. Since Bayesian theory is capable of revealing rational perceptual decision, we chose MoG as the supervised learning model. A cluster of Gaussians were applied on data of each class. We have proved that our representations do lead to similar representations between unsupervised and supervised learning, by systematically investigating the representations on four cognitive tasks: phoneme recognition, gender detection, speaker identity recognition and age detection. Phoneme, gender, speaker identity can be effortless recognized by humans. However age is also predicted from speech features corresponding to human ability to guess the speakers' age within a range. To test whether representations from unsupervised learning lead to similar errors in prediction of four speech-based tasks as in humans, we made investigation in a stepwise manner: from classification error rates, to sample-to-sample errors, and even the posterior probability level. High level correlation and consistency between two classifications has been found in all cognitive tasks.

All in all, our findings of COCA of speech signals are promising. The statistical regularities at multiple time scales corresponding to phoneme, gender, speaker identity and age have been revealed. Moreover the results indeed indicated the consistency of statistical regularities (unsupervised learning) and human cognitive processes (supervised learning of human labels). All these findings served as evidence to our speculation that '*ICA is employed by human brain in higher level cognitive activities*'.

References

Ahrendt, P., Meng, A., & Larsen, J. (2004). Decision time horizon for music genre classification using short time features. In *Proc. eusipco* (p. 1293-1296).

Arenas-Garca, J., Meng, A., Petersen, K. B., Schiler, T. L., Hansen, L. K., & Larsen, J. (2007). Unveiling music structure via plsa similarity fusion. In *Proc. ieee international workshop on machine learning for signal processing* (pp. 419–424).

Barlow, H. (1989). Unsupervised learning. *Neural Computation*, *1*, 295–311.

Bell, A. J., & Sejnowski, T. J. (1997). The 'independent components' of natural scenes are edge filters. *Vision Research*, *37*, 3327–3338.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. OXFORD University Press.

Blumstein, S. E., & Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *The Journal of the Acoustical Society of America*, *66*, 1001–1017.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing, Special issue on Higher-Order Statistics*, *36(3)*, 287-314.

Damper, R. I. (1998). Self-learning and self-organization as tools for speech research. *Behavioral and brain sciences*, *21*, 262–263.

Deerwester, S. C., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*, 391–407.

Deller, J. R., Hansen, J. H., & Proakis, J. G. (2000). *Discrete time processing of speech signals*. IEEE Press Marketing.

Doi, E., Inui, T., Lee, T. W., Wachtler, T., & Sejnowski, T. J. (2003). Spatiochromatic receptive field properties derived from information-theoretic analyses of cone mosaic responses to natural scenes. *Neural Comput.*, *15*, 397–417.

Dusan, S., & Rabiner, L. (2005). Can automatic speech recognition learn more from human speech perception. In *Proc. of the 3rd romanian academy conference on speech technology and human-computer dialogue* (pp. 21–36).

Feldman, J. (2004). *Bayes and the simplicity principle in perception* (Tech. Rep. No. 80). Rutgers University, Department of Psychoogy, Center for Cognitive Science.

Feng, L., & Hansen, L. K. (2005). On low level cognitive components of speech. In *Proc. international conference on computational intelligence for modelling* (Vol. 2, pp. 852–857).

Feng, L., & Hansen, L. K. (2006). Phonemes as short time cognitive components. In *Proc. icassp* (Vol. 5, p. 869-872).

Feng, L., & Hansen, L. K. (2007). Cognitive components of speech at different time scales. In *Proc. cogsci* (p. 983-988).

Feng, L., & Hansen, L. K. (2008). On phonemes as cognitive components of speech. In *Proc. iapr workshop on cognitive information processing*.

Field, D. J. (1994). What is the goal of sensory coding? *Neural Computation*, *6*, 559–601.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., & Dahlgren, N. L. (1993). The darpa timit acoustic phonetic continuous speech corpus cdrom. In *Nist order number pb91-100354*.

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, *47*, 103-138.

Hansen, L. K., Ahrendt, P., & Larsen, J. (2005). Towards cognitive component analysis. In *Akrr'05 -international and interdisciplinary conference on adaptive knowledge representation and reasoning*.

Hateren, J. H. van, & Ruderman, D. L. (1998). Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. In *Proc. biological sciences* (Vol. 265, pp. 2315–2320).

Hoyer, P., & Hyvrinen, A. (2000). Independent component analysis applied to feature extraction from colour and stereo images. *Network: Comput. Neural Syst.*, *11*, 191–210.

Hyvrinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. John Wiley-Interscience Publication.

Kintsch, W. (2001). Predication. *Cognitive Science*, *25*, 173–202.

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, *5*, 356-363.

Mather, G. (2006). *Foundations of perception*. Psychology Press.

Nielsen, A. B., Sigurdsson, S., Hansen, L. K., & Arenas-Garca, J. (2007). On the relevance of spectral features for instrument classification. In *Proc. icassp* (Vol. 5, p. 485-488).

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*, 607-609.

Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, *14(4)*, 481-487.

Olshausen, B. A., & O'Connor, K. N. (2002). A new window on sound. *Nature Neuroscience*, *5*, 292-294.

Pearlmutter, B. A., & Hinton, G. E. (1986). G-maximization: an unsupervised learning procedure for discovering regularities. In *Proc. aip conf. neural networks comp.* (p. 333-338).

Reisberg, D. (2006). *Cognition: Exploring the science of the mind*. W.W.Norton & Company.

Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using gaussian mixture models. *IEEE Trans. on Speech and Audio Processing*, *3(1)*, 72-83.

Salton, G. (1989). *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Addison-Wesley.

Slaney, M. (2002). Mixtures of probability experts for audio retrieval and indexing. In *Proc. ieee international conference on multimedia and expo* (p. 345-348).

Sussman, H. M., Fruchter, D., Hillbert, J., & Sirosh, J. (1998). Linear correlations in the speech signal: The orderly output constraint. *Behavioral and brain sciences*, *21*, 241–299.

Appendix I

Vocal Segment Classification in Popular Music

This article is accepted for publication in *Proc. the Ninth International Conference on Music Information Retrieval* 2008, pp 121-126, with the same title. Authors are Ling Feng, Andreas Brinch Nielsen and Lars Kai Hansen. It is also available as IMM publication database with number imm5654.

VOCAL SEGMENT CLASSIFICATION IN POPULAR MUSIC

ABSTRACT

This paper explores the vocal and non-vocal music classification problem within popular songs. A newly built labeled database covering 147 popular songs is announced. It is designed for classifying signals from $1sec$ time windows. Features are selected for this particular task, in order to capture both the temporal correlations and the dependencies among the feature dimensions. We systematically study the performance of a set of classifiers, including linear regression, generalized linear model, Gaussian mixture model, reduced kernel orthonormalized partial least squares and K-means on cross-validated training and test setup. The database is divided in two different ways: with/without artist overlap between training and test sets, so as to study the so called 'artist effect'. The performance and results are analyzed in depth: from error rates to sample-to-sample error correlation. A voting scheme is proposed to enhance the performance under certain conditions.

1 INTRODUCTION

The wide availability of digital music has increased the interest in music information retrieval, and in particular in features of music and of music meta-data, that could be used for better indexing and search. High-level musical features aimed at better indexing comprise, e.g., music instrument detection and separation [13], automatic transcription of music [8], melody detection [2], musical genre classification [10], sound source separation [18], singer recognition [16], and vocal detection [4]. While the latter obviously is of interest for music indexing, it has shown to be a surprisingly hard problem. In this paper we will pursue two objectives in relation to vocal/non-vocal music classification. We will investigate a multi-classifier system, and we will publish a new labeled database that can hopefully stimulate further research in the area.

While almost all musical genres are represented in digital forms, naturally popular music is most widely distributed, and in this paper we focus solely on popular music. It is not clear that the classification problem can be generalized between genres, but this is a problem we will investigate in later work.

Singing voice segmentation research started less than a decade ago. Berenzweig and Ellis attempted to locate the vocal line from music using a multi-layer perceptron speech model, trained to discriminate 54 phone classes, as the first step for lyric recognition [4]. However, even though singing and speech share certain similarities, the singing process involves the rapid acoustic variation, which makes it statistically different from normal speech. Such differences may lie in the phonetic and timing modification to follow the tune of the background music, and the usage of words or phrases in lyrics and their sequences. Their work was inspired by [15] and [19], where the task was to distinguish speech and music signals within the "music-speech" corpus: 240 $15s$ extracts collected 'at random' from the radio. A set of features have been designed specifically for speech/music discrimination, and they are capable of measuring the conceptually distinct properties of both classes.

Lyrics recognition can be one of a variety of uses for vocal segmentation. By matching the word transcriptions, it is applicable to search for different versions of the same song. Moreover, accurate singing detection could be potential for online lyrics display by automatically aligning the singing pieces with the known lyrics available on the Internet. Singer recognition of music recordings has later received more attention, and has become one of the popular research topics within MIR. In early work of singer recognition, techniques were borrowed from speaker recognition. A Gaussian Mixture Model (GMM) was applied based on Mel-frequency Cepstral Coefficients (MFCC) to detect singer identity [20]. As briefly introduced, singing voices are different from the conventional speech in terms of time-frequency features; and vocal and non-vocal features have differences w.r.t. spectral distribution. Hence the performance of a singer recognition system has been investigated using the unsegmented music piece, the vocal segments, and the non-vocal ones in [5]. 15% improvement has been achieved by only using the vocal segments, compared to the baseline of the system trained on the unsegmented music signals; and the performance became 23% worse when only non-vocal segments were used. It demonstrated that the vocal segments are the primary source for recognizing singers. Later, work on automatic singer recognition took vocal segmentation as the first step to enhance the system performance, e.g. [16].

Loosely speaking, vocal segmentation has two forms. One is to deal with a continuous music stream, and the locations of the singing voice have to be detected as well as classified, one example is [4]. The second one is to pre-segment the signals into windows, and the task is only to classify these segments into two classes. Our work follows the second line, in order to build models based on our in-house Pop

music database. A detailed description of the database will be presented in section 4. The voice is only segmented in the time domain, instead of the frequency domain, meaning the resulting vocal segments will still be a mixture of singing voices and instrumental background. Here we will cast the vocal segments detection in its simplest form, i.e. as a binary classification problem: one class represents signals with singing voices (with or without background music); the other purely instrumental segments, which we call accompaniment.

In this paper we study this problem from a different angle. Several classifiers are invoked, and the individual performance (errors and error rates) is inspected. To enhance performance, we study the possibility of sample-to-sample cross-classifier voting, where the outputs of several classifiers are merged to give a single prediction. The paper is organized as follows. Section 2 explains the selection of features. Classification frameworks are covered by section 3. With the purpose of announcing the Pop music database, we introduce the database design in section 4. In section 5, the experiments are described in depth, and the performance characteristics are presented. At last, section 6 concludes the current work.

2 ACOUSTIC FEATURES

2.1 Mel-Frequency Cepstral Coefficients

MFCCs are well-known in the speech and speaker recognition society. They are designed as perceptually weighted cepstral coefficients, since the mel-frequency warping emulates human sound perception. MFCCs share two aspects with the human auditory system: A logarithmic dependence on signal power and a simple bandwidth-to-center frequency scaling so that the frequency resolution is better at lower frequencies. MFCCs have recently shown their applicability in music signal processing realm, e.g. [1] for music genre classification, [16] and [5] for singer recognition, and [14] for vocal segmentation, and many more exist.

The features are basically extracted from short time scales, e.g. $20ms$, due to the stationarity of music signals. To process windows at longer time scales, temporal feature integration is needed. Features at different time scales may contain different information. A small frame size may result in a noisy estimation; and a long frame size may cover multiple sounds (phonemes) and fail to capture the appropriate information.

2.2 Multivariate AR

During the course of searching for appropriate features, researchers have realized that the systems performance can be improved by combining short-time frame-level features into clip-level features. Feature integration is one of the methods to form a long-time feature, in order to capture the discriminative information and characterize how frame-level features change over longer time periods for a certain task. Often the mean and variance of several short-time features are extracted as the clip-level features [17], using multivariate Gaussian model or a mixture of them. However, both the mean-variance and mean-covariance model fail to capture the temporal correlations. A frequency band approach has been proposed in [9], and the energy of the features was summarized into 4 frequency bands. Even though this method can represent temporal development, it does not model the feature correlations.

The multivariate autoregressive model (MAR) was recently introduced to music genre classification [11], and a detailed comparison of different temporal feature integration methods was reported. The superiority of MAR being able to capture both the temporal correlations and the dependencies among the feature dimensions, put this method into a league by itself. We adapt this model in the feature extraction phase on top of the short-time MFCCs. Here, a brief description of MAR will be given, for detail, see [11].

Assume the short-time MFCC at time t is denoted as $\mathbf{x}_t$, which is extracted from a short period of stationary signals. The MAR can be stated as,

$$\mathbf{x}_t = \sum_{p=1}^{P} \mathbf{A}_p \mathbf{x}_{t-p} + \mathbf{u}_t, \tag{1}$$

where $\mathbf{u}_t$ is the Gaussian noise $\mathcal{N}(\mathbf{v}, \mathbf{\Sigma})$, assumed i.i.d. $\mathbf{A}_p$ is the coefficients matrix for order p. P indicates the order of the multivariate auto regressive model, meaning that $\mathbf{x}_t$ is predicted from the previous P short-time features. It is worth to mention that the mean of the MFFCs $\mathbf{m}$ is related to the mean of the noise $\mathbf{v}$ in the following way,

$$\mathbf{m} = (\mathbf{I} - \sum_{p=1}^{P} \mathbf{A}_p)^{-1} \mathbf{v} \tag{2}$$

3 CLASSIFICATION FRAMEWORKS

We have examined a number of classifiers: linear regression model (LR), generalized linear model (GLM), Gaussian mixture model (GMM), reduced kernel orthonormalized partial least squares (rKOPLS) and K-means.

As the problem is a binary task, only a single dimension is needed for linear regression, and the labels are coded as ± 1. The model is $l_n = \mathbf{w}^{\mathbf{T}}\mathbf{y}$. A 1 is added to the feature vector to model offset. Least squares is used as the cost function for training, and the minimum solution is the pseudo inverse. The prediction is made based on the *sign* of the output: we tag the sample as a vocal segment if the output is greater than zero; and as a non-vocal segment otherwise.

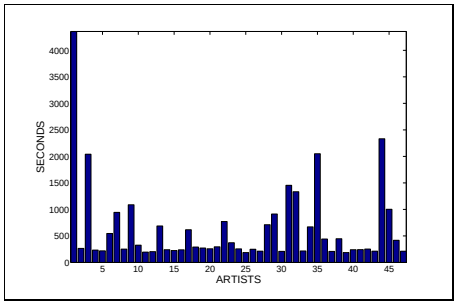

Figure 1. Distribution of Pop music among artists

Generalized linear model relates a linear function of the inputs, through a link function to the mean of an exponential family function, $g(\mu) = \mathbf{w}^T\mathbf{x}$. In our case we use the *softmax* link function, $\mu_i = \frac{e^{\mathbf{w}_i^T\mathbf{x}_i}}{\sum_j e^{\mathbf{w}_j^T\mathbf{x}_j}}$. $\mathbf{w}$ is found using iterative reweighted least squares [12].

GMM as one of the Bayesian classifiers, assumes a known probabilistic density distribution for each class. Hence we model data from each class as a group of Gaussian clusters. The parameters are estimated from training sets via the standard Expectation-Maximization (EM) algorithm. For simplicity, we assume the covariance matrices to be diagonal. Note that although features are independent within each mixture component due to the diagonal covariance matrix, the mixture model does not factorize over features. The diagonal covariance constraint posits the axes of the resulting Gaussian clusters parallel to the axes of the feature space. Observations are assigned to the class having the maximum *posterior* probability.

Any classification problem is solvable by a linear classifier if the data is projected into a high enough dimensional space (possibly infinite). To work in an infinite dimensional space is impossible, and kernel methods solve the problem by using inner products, which can be computed in the original space. Relevant features are found using orthonormalized partial least squares in kernel space. Then a linear classifier is trained and used for prediction. In the reduced form, rKOPLS [3] is able to handle large data sets, by only using a selection of the input samples to compute the relevant features, however all dimensions are used for the linear classifier, so this is not equal to a reduction of the training set.

K-means uses K clusters to model the distribution of each class. The optimization is done by assigning data points to the closest cluster centroid, and then updating the cluster centroid as the mean of the assigned data points. This is done iteratively, and minimizes the overall distances to cluster centroids. Optimization is very dependent on the initial centroids, and training should be repeated a number of times. Prediction is done by assigning a data point to the class of the closest cluster centroid.

	Error Rates
LR	19.03±2.25 %
GLM	18.46±2.02 %
GMM	23.27±2.54 %
rKOPLS	22.62±1.85 %
K-means	25.13±2.11 %

Table 1. The average error rates (mean ± standard deviation) of 5 classifiers on test sets.

4 DATABASE

The database used in the experiments is our recently built in-house database for vocal and non-vocal segments classification purpose. Due to the complexity of music signals and the dramatic variations of music, in the preliminary stage of the research, we focus only on one music genre: the popular music. Even within one music genre, Berenzweig, Ellis and Lawrence have pointed out the 'Album Effect'. That is songs from one album tend to have similarities w.r.t. audio production techniques, stylistic themes and instrumentation, etc [5].

This database contains 147 Pop mp3s: with 141 singing songs and 6 pure accompaniment songs. The 6 accompaniment songs are not the accompaniment of any of the other singing songs. The music in total lasts $8h\ 40min\ 2sec$. All songs are sampled at 44.1 kHz. Two channels are averaged, and segmentation is based on the mean. Songs are manually segmented into $1sec$ segments without overlap, and are annotated second-by-second. The labeling is based on the following strategy: if the major part of this $1sec$ music piece is singing voice, it is tagged as vocal segment; otherwise non-vocal segment. We believe that the long-term acoustic features are more capable of differentiating singing voice, and $1sec$ seems to be a reasonable choice based on [14]. Furthermore labeling signals at this time scale is not only more accurate, but also less expensive.

Usually the average partition of vocal/non-vocal in Pop music is about $70\%/30\%$. Around 28% of the 141 singing songs is non-vocal music in the collection of this database. Forty-seven artists/groups are covered. By artists in Pop music we mean the performers (singers) or bands instead of composers. The distribution of songs among artists is not even, and Figure 1 gives the total number of windows (seconds) each artist contributes.

5 EXPERIMENTS AND RESULTS

We have used a set of features extracted from the music database. First, we extracted the first 6 original MFCCs over

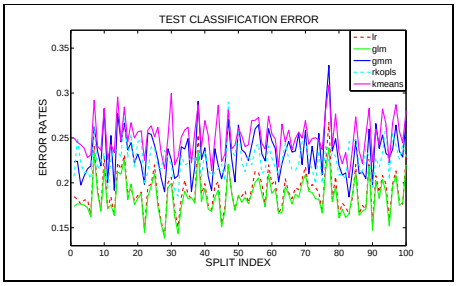

Figure 2. Classification error rates as a function of splits of five classifiers on test sets.

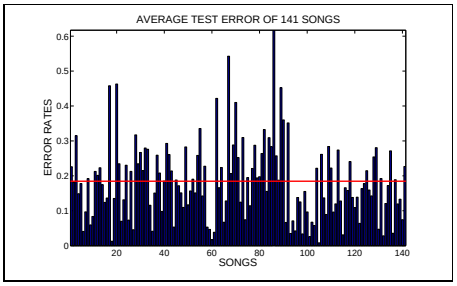

Figure 3. Test classification error rates for individual songs by GLM model. The dash line gives the average error rates of the 100-split cross-validation.

a $20ms$ frame hopped every $10ms$. The 0^{th} MFCC representing the log-energy was computed as well. The means were calculated on signals covering $1sec$ in time. MAR were afterwards computed on top of the first 6 MFCCs with $P = 3$, and we ended up with a 6-by-18 $\mathbf{A}_p$ matrix, a 1-by-6 vector $\mathbf{v}$ and a 6-by-6 covariance matrix Σ. Since Σ is symmetric, the repetitions were discarded. $\mathbf{A}_p$, $\mathbf{v}$ and Σ all together form a 135-dimensional feature set. All in all, for $1sec$ music signal we concatenated 135-d MAR, the means of both 0^{th} and 6 original MFCCs to form a 142-d feature vector.

5.1 Data Dependency and Song Variation

We used one type of cross-validation, namely holdout validation, to evaluate the performance of the classification frameworks. To represent the breadth of available signals in the database, we kept 117 songs with the 6 accompaniment songs to train the models, and the remaining 30 to test. We randomly split the database 100 times and evaluated each classifier based on the aggregate average. In this way we eliminated the data set dependencies, due to the possible similarities between certain songs. The random splitting regarded a song as one unit, therefore there was no overlap song-wise in the training and test set. On the other hand artist overlap did exist. The models were trained and test set errors were calculated for each split. The GLM model from the Netlab toolbox was used with *softmax* activation function on the output, and the model was trained using iterative reweighted least squares. As to GMM, we used the generalizable gaussian mixture model introduced in [7], where the mean and variance of GMM are updated with separate subsets of the data. We fixed the number of Gaussian mixtures as 4 for non-vocal model, and 8 for vocal model. For rKOPLS, we randomly chose 1000 windows from the training set to calculate the feature projections. The average error rates of the five classification algorithms are summarized in Table 1.

A bit surprisingly the performance is significantly better

	LR	GLM	GMM	rKOPLS	K-means
LR	1.0000	0.9603	0.8203	0.8040	0.8110
GLM	0.9603	1.0000	0.8141	0.8266	0.8091
GMM	0.8203	0.8141	1.0000	0.7309	0.7745
rKOPLS	0.8040	0.8266	0.7309	1.0000	0.7568
K-means	0.8110	0.8091	0.7745	0.7568	1.0000

Table 2. A matrix of the degree of matching.

for the linear models. We show the performance of the chosen classifiers as a function of splits in Figure 2. Each curve represents one classifier, and the trial-by-trial difference is quite striking. It proved our assumption that the classification performance depends heavily on the data sets, and the misclassification varies between 13.8% and 23.9% for the best model (GLM). We envision that there is significant variation in the data set, and the characteristics of some songs may be distinguishing to the others. To test the hypothesis, we studied the performance on individual songs. Figure 3 presents the average classification errors of each song predicted by the best model: GLM, and the inter-song variation is obviously revealed: for some songs it is easy to distinguish the voice and music segments; and some songs are hard to classify.

5.2 Correlation Between Classifiers and Voting

While observing the classification variation among data splits in Figure 2, we also noticed that even though classification performance is different from classifier to classifier, the tendency of these five curves does share some similarity. Here we first carefully studied the pair-to-pair performance correlation between the classification algorithms. In Table 2 the degree of matching is reported: 1 refers to perfect match; 0 to no match. It seems that the two linear classifiers have a very high degree of matching, which means that little will be gained by combining these two.

The simplest way of combining classification results is

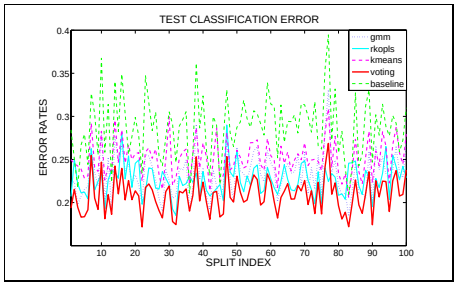

Figure 4. Voting results. It gives the voting performance among GMM, rKOPLS and K-means. The light dash line shows the baseline of random guessing for each data split.

by majority voting, meaning that the class with the most votes is chosen as the output. The voting has been done crossing all five classifiers, unfortunately the average voting results (error rates) on the test sets was 18.62%, which is slightly worse than the best individual classifier. The reason seems to be that even though the other classifiers are not so correlated with the linear ones, the miss classification rate is too high to improve performance.

However voting does help enhance the performance, if it performs among not so correlated classification results. Figure 4 demonstrates the sample-to-sample majority voting among three classifiers: GMM, rKOPLS and K-means. The similar tendency was preserved in the voting results, and there were only 10 data splits out of 100, where the voting results were worse than the best ones among these three. The average performance of voting on test sets was $20.90 \pm 2.02\%$.

Here we will elaborate on the performance on individual songs, by looking at the predicted labels from each classifier and voting predictions. Figure 5 demonstrates how voting works, and how the prediction results correlate. Two songs: 'Do You Know What You Want' by M2M, and 'A Thousand Times' by Sophie Zelmani, have been chosen to illustrate the 'good' and 'bad' cases, i.e. when voting helps and fails. Vocal segments are tagged with '1', and '0' for non-vocal ones. The ground truth is given as a reference. The voting was carried out among GMM, rKOPLS and K-means, and their predictions are shown. If the classifiers make mistakes in a similar pattern, the voting cannot recover the wrong predictions, e.g. area B. If the predictions are not correlated to a high degree voting helps, e.g. area A.

Moreover, we noticed that it is very likely for classifiers to make wrong predictions in the transition sections, meaning the changing from vocal to non-vocal parts, and vice versa. We found this is reasonable comparing with manual labels by different persons, shown in Figure 6. The song was labeled carefully by both people, the absence of mind or

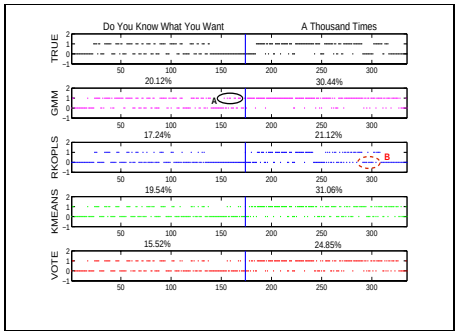

Figure 5. Sample-to-sample errors and voting results. Two songs represent the 'good' and 'bad' voting cases. Individual error rates for each classifier and voting results are given. Two areas marked A & B indicate the scenarios when voting helps and fails.

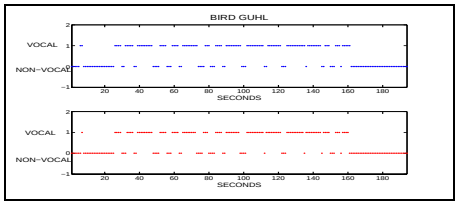

Figure 6. Two manual label results of the same song: 'Bird Guhl'. It is obvious that the disagreement only appears in the transition parts.

guessing should not be a concern. The mismatch indicates the perception or judging difference, and it only happens in the transition parts. The total mismatch is about 3% for this particular song: 'Bird Guhl' by Antony and the Johnsons.

5.3 'Artist Effect'

In previous experiments, we randomly selected songs to form training and test sets, hence the same artist may appear in both sets. Taking the previous results as a baseline, we studied the 'artist effect' in this classification problem. We tried to keep the size of test sets the same as before, and carefully selected around 30 songs in order to avoid artist overlap for each split, and formed 100 splits. Table 3 summarizes the average error rates for 4 classifiers. The average results are a little worse than the previous ones, and they also have bigger variance along the splits. Therefore we speculate that artists do have some influence in vocal/non-vocal music classification, and the influence may be caused by different styles, and models trained on particular styles are hard to be generalized to other styles.

	Error Rates
LR	20.52±3.5 %
GLM	19.82±2.81 %
GMM	24.50±2.99 %
rKOPLS	24.60±3.14 %

Table 3. Averaged test error rates of 4 classifiers on cross-validation without artist overlap.

6 CONCLUSION AND DISCUSSION

We have investigated the voal/non-vocal popular music classification. Experiments were carried out on our database, containing 147 popular songs. To be in line with the label set, the classifiers were trained based on features at $1sec$ time scale. We have employed 142-d acoustic features, consisting MFCCs and MAR, to measure the distinct properties of vocal and non-vocal music. Five classifiers have been invoked: LR, GLM, GMM, rKOPLS and K-means.

We cross-validated the whole database, and measured the aggregate average to eliminate the data set dependency. The GLM outperformed all the others, and provided us with 18.46% error rate on the baseline of 28%. The performance has great variation among data splits and songs, indicating the variability of popular songs. The correlations among classifiers have been investigated, and the proposed voting scheme helped among less correlated classifiers. Finally we looked into the 'artist effect', and it did degrade the classification accuracy a bit by separating artists in training and test sets. All in all vocal/non-vocal music classification was found to be a difficult problem, and it depends heavily on the music itself. Maybe classification within similar song styles can improve the performance.

7 REFERENCES

[1] Ahrendt, P., Meng, A. and Larsen, J. "Decision time horizon for music genre classification using short time features", *Proceedings of EUSIPCO*, pp. 1293-1296, 2004.

[2] Akeroyd, M. A., Moore, B. C. J. and Moore, G. A. "Melody recognition using three types of dichotic-pitch stimulus", *The Journal of the Acoustical Society of America*, vol. 110, Issue 3, pp. 1498-1504, 2001.

[3] Arenas-García, J., Petersen, K.B., Hansen, L.K. "Sparse Kernel Orthonormalized PLS for feature extraction in large data sets", *Proceedings of NIPS*, pp. 33-40, MIT Press, Cambridge, MA, 2007.

[4] Berenzweig, A. L., Ellis, D. P. W., and Lawrence, S. "Locating Singing Voice Segments Within Music Signals", *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Aoustics*, New Paltz, NY, 2001.

[5] Berenzweig, A. L., Ellis, D. P. W., and Lawrence, S. "Using Voice Segments to Improve Artist Classification of Music", *Proceedings of the International Conference on Virtual, Synthetic and Entertainment Audio*, Espoo, Finland, 2002.

[6] Eronen, A. "Automatic musical instrument recognition". Master Thesis, Tampere University of Technology, 2001.

[7] Hansen, L. K., Sigurdsson, S., Kolenda, T., Nielsen, F. ., Kjems, U., Larsen, J. "Modeling text with generalizable gaussian mixtures", *Proceedings of ICASSP*, vol. 4, pp. 3494-3497, 2000

[8] Heln, M. and Virtanen, T. "Separation of Drums From Polyphonic Music Using Non-Negative Matrix Factorization and Support Vector Machine", *Proceedings of EUSIPCO*, Antalya, Turkey, 2005.

[9] McKinney, M. F. and Breebart, J. "Features for audio and music classification", *Proceedings of ISMIR*, Baltimore, Maryland (USA), pp.151-158, 2003.

[10] Meng, A., Shawe-Taylor J., "An Investigation of Feature Models for Music Genre Classification using the Support Vector Classifier", *Proceedings of ISMIR*, London, UK, pp. 604-609, 2005.

[11] Meng, A., Ahrendt, P., Larsen, J. and Hansen, L. K. "Temporal Feature Integration for Music Genre Classification", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15(5), pp. 1654-1664, 2007.

[12] Nabney, I., Bishop, C. "Netlab neural network software", ver. 3.2, 2001

[13] Nielsen, A. B., Sigurdsson, S., Hansen, L. K., and Arenas-Garca, J. "On the relevance of spectral features for instrument classification", *Proceedings of ICASSP*, Honolulu, Hawaii, vol. 5, pp. 485-488, 2007.

[14] Nwe, T. L. and Wang, Y. "Automatic Detection of Vocal Segments in Popular Songs" *Proceedings of the ISMIR*, Barcelona, Spain, 2004.

[15] Scheirer, E. and Slaney, M. "Construction and Evaluation fo A Robust Multifeature Speech/Music Discriminator", *Proceedings of ICASSP*, Munich, 1997.

[16] Tsai W.-H. and Wang, H.-M. "Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals", *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 330–341, 2006.

[17] Tzanetakis, G. "Manipulation, analysis and retrieval systems for audio signal", *Ph.D. dissertation* , Faculty of Princeton University, Department of Computer Science, 2002

[18] Virtanen, T. "Separation of Sound Sources by Convolutive Sparse Coding", *Proceedings of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing, SAPA*, 2004.

[19] Williams, G. and Ellis, D. P. W. "Speech/Music Discrimination Based on Posterior Probability Features", *Proceedings of Eurospeech*, Budapest, 1999.

[20] Zhang, T. "Automatic singer identification", *Proceedings of ICME*, Baltimore, 2003.

Bibliography

[1] L. A. Adamic and B. A. Huberman. Power-law distribution of the world wide web. *Science*, 287, 2000.

[2] P. Ahrendt, A. Meng, and J. Larsen. Decision time horizon for music genre classification using short time features. In *Proc. EUSIPCO*, pages 1293–1296, 2004.

[3] M. A. Akeroyd, B. C. J. Moore, and G. A. Moore. Melody recognition using three types of dichotic-pitch stimulus. *The Journal of the Acoustical Society of America*, 110(3):1498–1504, 2001.

[4] R. Albert, A.L. Barabsi, and H. Jeong. Scale-free characteristics of random networks: The topology of the world wide web. *Physica A*, 281:69–77, 2000.

[5] R. Albert, H. Jeong, and A.L. Barabási. Diameter of the world wide web. *Nature*, 401(130):130–131, 1999.

[6] J. Arenas-García, A. Meng, K. B. Petersen, T. L. Schiøler, L. K. Hansen, and J. Larsen. Unveiling music structure via plsa similarity fusion. In *Proc. IEEE International Workshop on Machine Learning for Signal Processing*, pages 419–424, 2007.

[7] H.B. Barlow. Unsupervised learning. *Neural Computation*, 1:295–311, 1989.

[8] A. J. Bell and T. J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37:3327–3338, 1997.

[9] C. M. Bishop. *Neural Networks for Pattern Recognition*. OXFORD University Press, 1995.

[10] S. E. Blumstein and K. N. Stevens. Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *The Journal of the Acoustical Society of America*, 66:1001–1017, 1979.

[11] J. M. Bower and L. M. Parsons. Rethinking the 'lesser brain'. *Scientific American Magazine*, 289:50–57, 2003.

[12] P. Comon. Independent component analysis, a new concept? *Signal Processing, Special issue on Higher-Order Statistics*, 36(3):287–314, 1994.

[13] L. P. Cordella, P. Foggia, C. Sansone, and M. Vento. A real-time text-independent speaker identification system. In *Proc. ICIAP*, pages 632–637, 2003.

[14] R. I. Damper. Self-learning and self-organization as tools for speech research. *Behavioral and brain sciences*, 21:262–263, 1998.

[15] S. C. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

[16] J. R. Deller, J. H. Hansen, and J. G. Proakis. *Discrete Time Processing of Speech Signals*. IEEE Press Marketing, 2000.

[17] P. Domingos and M. Pazzani. Beyond independence: Conditions for the optimality of the simple bayesian classifier. *Machine Learning*, 29:103–130, 1997.

[18] D. L. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts. In *Proc. NIPS*, 2003.

[19] S. Dusan and L. Rabiner. Can automatic speech recognition learn more from human speech perception. In *Proc. of the 3rd Romanian Academy Conference on Speech Technology and Human-Computer Dialogue*, pages 21–36, 2005.

[20] M. Dyrholm, S. Makeig, and L. K. Hansen. Model selection for convolutive ica with an application to spatio-temporal analysis of eeg. *Neural Computation*, 19(4):934–955, 2007.

[21] N. M. J. Edelstyn and F. Oyebode. A review of the phenomenology and cognitive neuropsychological origins of the capgras syndrome. *The International Journal of Geriatric Psychiatry*, 14:48–59, 1999.

[22] Jacob Feldman. Bayes and the simplicity principle in perception. Technical Report 80, Rutgers University, Department of Psychoogy, Center for Cognitive Science, 2004.

[23] L. Feng and L. K. Hansen. Elsdsr english language speech database for speaker recognition, http://www2.imm.dtu.dk/ ∼ lf/ELSDSR.htm. 2004.

[24] D. J. Field. What is the goal of sensory coding? *Neural Computation*, 6:559–601, 1994.

[25] E. Frank, L. Trigg, G. Holmes, and I. H. Witten. Naive bayes for regression. *Machine Learning*, 41(1):5–26, 2000.

[26] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. Statistical semantics: Analysis of the potential performance of key-word information systems. *Bell System Technical Journal*, 62(6):1753–1806, 1983.

[27] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. The darpa timit acoustic phonetic continuous speech corpus cdrom. In *NIST order number PB91-100354*. 1993.

[28] Z. Ghahramani and G. E. Hinton. The em algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, University of Toronto, Department of Computer Science, 6 King's College Road, Toronto, Canada M5S 1A4, 1996.

[29] B. R. Glasberg and B. C. J. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47:103–138, 1990.

[30] L. K. Hansen, P. Ahrendt, and J. Larsen. Towards cognitive component analysis. In *AKRR05-International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, 2005.

[31] L. K. Hansen, J. Larsen, and T. Kolenda. On independent component analysis for multimedia signals. pages 175–199, 2000.

[32] L. K. Hansen, J. Larsen, and T. Kolenda. Blind detection of independent dynamic components. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 3197–3200, 2001.

[33] S. Haykin and Z. Chen. The cocktail party problem. *Neural Comp.*, 17:1875–1902, 2005.

[34] M. Helén and T. Virtanen. Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In *Proc. EUSIPCO*, 2005.

[35] P. Hoyer and A. Hyvärinen. Independent component analysis applied to feature extraction from colour and stereo images. *Network: Comput. Neural Syst.*, 11:191–210, 2000.

[36] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

[37] A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

[38] W. James. The principles of psychology. *New York: Henry Holt*, 1:403–404, 1890.

[39] H. G. Kim, E. Berdahl, N. Moreau, and T. Sikora. Speaker recognition using mpeg-7 descriptors. In *Proc. Eurospeech*, pages 489–492, 2003.

[40] W. Kintsch. Predication. *Cognitive Science*, 25:173–202, 2001.

[41] T. Kolenda, L. K. Hansen, and J. Larsen. Signal detection using ica: Application to chat room topic spotting. In *Proc. International Conference on Independent Component Analysis and Blind Source Separation*, pages 540–545, 2001.

[42] T. Kolenda, L. K. Hansen, J. Larsen, and O.Winther. Independent component analysis for understanding multimedia content. In *Proc. IEEE Workshop on Neural Networks for Signal Processing XII*, pages 757–766, 2002.

[43] I. Kononenko. *Comparison of Inductive and Naive Bayesian Learning Approaches to Automatic Knowledge Acquisition. Current Trends in Knowledge Acquisition.* IOS Press, 1990.

[44] J. Larsen, L.K. Hansen, T. Kolenda, and F.AA. Nielsen. Independent component analysis in multimedia modeling. In *Proc. International Symposion on Independent Component Analysis and Blind Source Separation*, pages 687–696, 2003.

[45] D. D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[46] D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference*, pages 556–562, 2001.

[47] I. Lee, T. Kim, and T. W. Lee. Fast fixed-point independent vector analysis algorithms for convolutive blind source separation. *Signal Processing*, 87(8):1859–1871, 2007.

[48] S. Lehmann, B. E. Lautrup, and A. D. Jackson. Citation networks in high energy physics. *Physical Review E*, 68:026113–026120, 2003.

[49] M. S. Lewicki. Efficient coding of natural sounds. *Nature Neuroscience*, 5:356–363, 2002.

[50] E. E. Loos, S. Anderson, D. H. Jr. Day, P. C. Jordan, and J. D. Wingate. Glossary of linguistic terms. In *SIL International, http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/index.htm.* 2004.

[51] W. G. Lycan. *Mind and Cognition: An Anthology, Second Edition*. Oxford: Basil Blackwell, 1999.

[52] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003. Available from http://www.inference.phy.cam.ac.uk/mackay/itila/.

[53] G. Mather. *Foundations of Perception*. Psychology Press, 2006.

[54] T. Matsui and S. Furui. Comparison of text-independent speaker recognition methods using vq-distortion and discrete/continuous hmms. In *Proc. ICASSP*, volume 2, pages 157–160, 1992.

[55] M. F. McKinney and J. Breebart. Features for audio and music classification. In *Proc. International Symposium on Music Information Retrieval*, pages 151–158, 2003.

[56] A. Meng, P. Ahrendt, and J. Larsen. Improving music genre classification by short-time feature integration. In *Proc. ICASSP*, volume 5, pages 497–500, 2005.

[57] A. Meng, P. Ahrendt, J. Larsen, and L. K. Hansen. Temporal feature integration for music genre classification. *IEEE Transactions on Audio, Speech, and Language Processing*, 15:1654–1664, 2007.

[58] A. Meng and J. Shawe-Taylor. An investigation of feature models for music genre classification using the support vector classifier. In *Proc. ISMIR*, pages 604–609, 2005.

[59] P. Moerland. Mixtures of latent variable models for density estimation and classication. Technical Report IDIAP-RR 25, IDIAP: Institut Dalle Molle d'Intelligence Artificielle Perceptive, Centre du Parc Av. des Prés-Beudin 20 Case Postale 592 CH-1920 Martigny Switzerland, 2000.

[60] B. C. J. Moore. *An Introduction of the Psychology of Hearing*. Academic Press, 2004.

[61] M. Mørup, K. H. Madsen, and Lars Kai Hansen. Shifted independent component analysis. In *Proc. ICA*, pages 89–96, 2007.

[62] H. A. Murthy, F. Beaufays, L. P. Heck, and M. Weintraub. Robust text-independent speaekr identification over telephone channels. *IEEE trans. on Speech and Audio Processing*, 7(5):554–568, 1999.

[63] A. B. Nielsen, S. Sigurdsson, L. K. Hansen, and J. Arenas-García. On the relevance of spectral features for instrument classification. In *Proc. ICASSP*, volume 5, pages 485–488, 2007.

[64] P. D. O'Grady and B. A. Pearlmutter. Hard-lost: Modified k-means for oriented lines. In *Proc. the Irish Signals and Systems Conference*, 2004.

[65] P. D. O'Grady and B. A. Pearlmutter. Soft-lost: Em on a mixture of oriented lines. In *Proc. ICA*, pages 430–436, 2004.

[66] B. A. Olshausen and D. J. Field. Emergence of simple cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[67] B. A. Olshausen and D. J. Field. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14(4):481–487, 2004.

[68] B. A. Olshausen and K. N. O'Connor. A new window on sound. *Nature Neuroscience*, 5:292–294, 2002.

[69] B. A. Pearlmutter and G. E. Hinton. G-maximization: an unsupervised learning procedure for discovering regularities. In *Proc. AIP Conference 151 on Neural Networks for Computing*, pages 333–338, 1986.

[70] S. Redner. Citation statistics from more than a century of physical review. *In APS March Meeting. American Physical Society*, 2005.

[71] D. Reisberg. *Cognition: Exploring the Science of the Mind.* W.W.Norton & Company, 2006.

[72] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using gaussian mixture models. *IEEE Trans. on Speech and Audio Processing*, 3(1):72–83, 1995.

[73] J. Rosca and A. Kofmehl. Cepstrum-like ica representations for text indepedetn speaker recognition. In *Proc. ICA*, pages 999–1004, 2003.

[74] G. Salton. *Automatic text processing: The transformation, analysis, and retrieval of information by computer.* Addison-Wesley, 1989.

[75] M. Slaney. Mixtures of probability experts for audio retrieval and indexing. In *Proc. IEEE International Conference on Multimedia and Expo*, pages 345–348, 2002.

[76] S. S. Stevens and J. Volkman. The relatino of pitch to frequency. *American Journal of Psychology*, 53:329, 1940.

[77] C. Tanprasert, C. Wutiwiwatchai, and S. Sae-tang. Text-dependent speaker identification using neural netwrok on distinctive thai tone marks. In *Proc. IJCNN International Joint Conference on Neural Network*, volume 5, pages 2950–2953, 1999.

[78] M. E. Tipping and C. M. Bishop. Mixture of probabilistic principal component analysers. *Neural Computation*, 11(2):443–482, 1999.

[79] D. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, 1985.

[80] W.-H. Tsai and H.-M. Wang. Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals. *IEEE Trans. on Audio, Speech, and Language Processing*, 14(1):330–341, 2006.

[81] G. Tzanetakis. Ph.d. dissertation: 'manipulation, analysis and retrieval systems for audio signals'. Technical Report CRG-TR-96-1, Faculty of Princeton University, Department of Computer Science, 6 King's College Road, Toronto, Canada M5S 1A4, 2002.

[82] T. Virtanen. Separation of sound sources by convolutive sparse coding. In *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing, SAPA*, 2004.

[83] J. Wagensberg. Complexity versus uncertainty: The question of staying alife. *Biology and philosophy*, 15:493–508, 2000.

[84] Y. Wang, Z. Liu, and J.C. Huang. Multimedia content analysis using both audio and visual clues. *IEEE Signal Processing Magazine*, 17(6):2–36, 2000.

[85] E. Wold, T. Blum, D. Keislar, and J. Wheaton. Content-based classification, search, and retrieval of audio. *IEEE Multimedia*, 3(3):27–36, 1996.

[86] Y. Zhang and J. Zhou. Audio segmentation based on multi-scale audio classification. In *Proc. ICASSP*, pages 349–352, 2004.

[87] F. Zheng, G. L. Zhang, and Z. J. Song. Comparison of different implementations of mfcc. *J. Computer Science & Technology*, 16(6):582–589, 2001.

[88] X. Zhu. Semi-supervised learning literature survey. Technical report, Computer Sciences TR 1530, University of Wisconsin Madison, 2007.